IMPROVING KNOWLEDGE ACCESSIBILITY ON THE WEB

– from Knowledge Base Augmentation to Search as Learning

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Ran Yu

from Xilinhot, Nei Mongol, China

Cologne, September 2019

from the institute for Informatik
at the Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisor: Prof. Dr. Stefan Dietze
Co-supervisor: Prof. Dr. Stefan Conrad
Co-supervisor: Prof. Dr. Claudia Hauff

Date of the oral examination: 13. February 2020

Statutory Declaration

I herewith formally declare that I have written the submitted dissertation independently. I did not use any outside support except for the quoted literature and other sources mentioned in the thesis.

I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content.

I am aware that the violation of this regulation will lead to failure of the thesis.

Place, Date Cologne, 16. Mar 2020 Signature 

ABSTRACT

The World Wide Web constitutes the largest collection of knowledge and is accessed by billions of users in their daily lives through applications such as search engines and smart assistants. However, most of the knowledge available on the Web is unstructured and is difficult for machines to process which leads to the lowered performance of such smart applications. Hence improving the accessibility of knowledge on the Web for machines is a prerequisite for improving the performance of such applications. Knowledge base as one of the most commonly used types of machine-readable knowledge resources, is inherently incomplete, particularly with respect to tail entities and properties. Improving the completeness and correctness of knowledge bases is one of the major challenges for improving the knowledge accessibility for machines.

Web search is one of the most ubiquitous online activities, commonly used to acquire new knowledge and to satisfy learning-related objectives. The importance of learning as an outcome of Web search has been recognized widely, leading to a variety of research at the intersection of information retrieval, human-computer interaction and learning-oriented sciences. Yet, there is a lack of understanding of the impact of Web search on a user's knowledge state. Understanding and automatically predicting the knowledge gain of users can be an important step forward if Web search engines that are currently optimized for relevance can be molded to better serve human learning needs.

In this thesis, we focus on improving the accessibility of knowledge on the Web for both machines and humans. We carried out comprehensive analysis of knowledge resources and learning related Web search sessions. Furthermore, we propose automated approaches to improve the completeness and correctness of knowledge bases and to allow search systems to understand human learning. To this end we make the following contributions as part of this thesis:

- *Knowledge Base Augmentation with Structured Web Markup.* As a complementary data source, embedded entity markup based on Microdata, RDFa, and Microformats have become prevalent on the Web and constitute an unprecedented source of data with significant potential to aid the task of knowledge base augmentation (KBA). RDF statements extracted from markup are fundamentally different from traditional knowledge graphs: entity descriptions are flat, facts are highly redundant and of varied quality, and, explicit links are missing despite a vast amount of coreferences. We present a novel approach which addresses these issues through a combination of entity matching and fusion techniques geared towards the specific challenges associated with Web markup. To ensure precise and non-redundant results, we follow a supervised learning ap-

proach based on a set of features considering aspects such as quality and relevance of entities, facts and their sources. We perform a thorough evaluation on a subset of the Web Data Commons dataset and show significant potential for augmenting existing knowledge bases. A comparison with existing data fusion baselines demonstrates the superior performance of our approach when applied to Web markup data.

- *Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web.* We present a study addressing the knowledge gain of users in informational search sessions. Using crowdsourcing, we recruited 500 distinct users and orchestrated real-world search sessions spanning 10 different topics and information needs. By using scientifically formulated knowledge tests we calibrated the knowledge of users before and after their search sessions, quantifying their knowledge gain. We investigated the impact of information needs on the search behavior and knowledge gain of users, revealing a significant effect of information need on user queries and navigational patterns, but no direct effect on the knowledge gain. Users on average exhibited a higher knowledge gain through search sessions pertaining to topics they were less familiar with.

- *Predicting User Knowledge Gain in Informational Search Sessions.* We introduce supervised models to predict a user's knowledge state and knowledge gain from features captured during a search session. Our supervised models utilise and derive a comprehensive set of features from the current state-of-the-art and compare the performance of a range of feature sets and feature selection strategies. Through our results, we demonstrate the ability to predict and classify the knowledge state and gain using features obtained during search sessions. Our models exhibit superior performance to an existing baseline in the knowledge state prediction task.

- *Topic-independent Modeling of User Knowledge in Informational Search Sessions.* Our previous investigation shows that it is possible to build supervised models to predict a user's knowledge gain and knowledge state from user interactions during a search session. However, the characteristics of the resources that a user interacts with have neither been sufficiently explored, nor exploited in this task. Hence, we further our exploration and introduce a novel set of resource-centric features and demonstrate their capacity to significantly improve supervised models for the task of predicting knowledge gain and knowledge state of users in Web search sessions. We make important contributions, given that reliable training data for such tasks is sparse and costly to obtain. More importantly, we introduce various feature selection strategies geared towards selecting a limited subset of effective and generalizable features.

The experimental result demonstrates that our approach improves the performance of knowledge prediction models on search sessions of unseen topics.

Keywords: *knowledge base augmentation, search as learning, user modeling, search log analysis*

ZUSAMMENFASSUNG

Das World Wide Web stellt die größte Sammlung menschlichen Wissens dar und wird von Milliarden von Nutzern in ihrem täglichen Leben über Anwendungen wie Suchmaschinen und intelligente Assistenten genutzt. Der größte Teil des im Web verfügbaren Wissens ist jedoch unstrukturiert und für Maschinen schwer zu verarbeiten, was zu einer geringeren Leistung solcher intelligenten Anwendungen führt. Die Verbesserung der Maschinenlesbarkeit von Wissen im Web ist daher eine Voraussetzung für die Verbesserung der Performance solcher Anwendungen. Wissensdatenbanken als eine der am häufigsten verwendeten Arten von maschinenlesbaren Ressourcen sind von Natur aus unvollständig, insbesondere in Bezug auf Tail-Entitäten und Eigenschaften. Die Verbesserung der Vollständigkeit und Korrektheit von Wissensdatenbanken ist eine der größten Herausforderungen bei der Verbesserung der Zugänglichkeit von Wissen für Maschinen.

Die Websuche ist eine der allgegenwärtigsten Online-Aktivitäten, die häufig genutzt wird, um neues Wissen zu erwerben und lernbezogene Ziele zu erreichen. Die Bedeutung von Lernen als Ergebnis der Websuche wurde allgemein anerkannt, was zu einer Vielzahl von Forschungsarbeiten an der Schnittstelle von Informationsbeschaffung, Mensch-Computer-Interaktion und lernorientierten Wissenschaften führte.

Dennoch fehlt es an Verständnis für die Auswirkungen der Websuche auf den Wissensstand eines Benutzers. Das Verstehen und die automatische Vorhersage des Wissenszuwachses der Nutzer kann ein wichtiger Schritt nach vorne sein, wenn Websuchmaschinen, die derzeit für die Relevanz optimiert sind, so gestaltet werden können, dass sie den menschlichen Lernergebnissen dienen.

In dieser Arbeit konzentrieren wir uns auf die Verbesserung der Zugänglichkeit von Wissen im Web für Maschinen und Menschen. Wir haben eine umfassende Analyse der Wissensressourcen und lernbezogene Suchvorgänge durchgeführt. Darüber hinaus schlagen wir automatisierte Ansätze vor, die die Vollständigkeit und Korrektheit der Wissensdatenbanken verbessern und es Suchsystemen ermöglichen, das Lernen der Benutzer zu verstehen. Zu diesem Zweck leisten wir im Rahmen dieser Arbeit die folgenden Beiträge:

- *Knowledge Base Augmentation mit Structured Web Markup.* Als ergänzende Datenquelle hat sich das Embedded Entity Markup auf Basis von Mikrodaten, RDFa und Mikroformaten im Web durchgesetzt und stellt eine beispiellose Datenquelle mit erheblichem Potenzial zur Unterstützung der Aufgabe der Wissensbasis-Augmentation (KBA) dar. RDF-Anweisungen, die aus Markup extrahiert werden, unterscheiden sich grundlegend

von traditionellen Wissensdiagrammen: Entitätsbeschreibungen sind flach, Fakten sind hoch redundant und von unterschiedlicher Qualität. Trotz einer Vielzahl von Co-Referenzen fehlen explizite Links.

Wir präsentieren einen neuartigen Ansatz, der diese Probleme durch eine Kombination von Entity-Matching und Fusionstechniken löst, die auf die spezifischen Herausforderungen im Zusammenhang mit Web-Markup zugeschnitten sind. Um präzise und nicht redundante Ergebnisse zu gewährleisten, verfolgen wir einen überwachten Lernansatz, der auf einer Reihe von Merkmalen basiert, die Aspekte wie Qualität und Relevanz von Einheiten, Fakten und deren Quellen berücksichtigen. Wir führen eine gründliche Evaluierung eines Teilsatzes des Web Data Commons Datensatzes durch und zeigen signifikantes Potenzial für die Erweiterung bestehender Wissensbestände. Ein Vergleich mit bestehenden Datenfusionsbasislinien zeigt eine überlegene Leistung unseres Ansatzes bei der Anwendung auf Web-Markup-Daten.

- *Analyse des Wissensvorsprungs von Benutzern in informativen Suchsitzungen im Web.* Wir stellen eine Studie vor, die sich mit dem Wissensgewinn der Nutzer bei der Informationssuche befasst. Mit Hilfe von Crowd-sourcing rekrutierten wir 500 verschiedene Benutzer und organisierten reale Suchsitzungen, die 10 verschiedene Themen und Informationsbedürfnisse abdeckten. Mit Hilfe wissenschaftlich formulierter Wissenstests kalibrieren wir das Wissen der Nutzer vor und nach ihrer Suche und quantifizieren ihren Erkenntnisgewinn. Wir untersuchten die Auswirkungen des Informationsbedarfs auf das Suchverhalten und den Wissensgewinn der Nutzer und zeigten einen signifikanten Einfluss des Informationsbedarfs auf Benutzeranfragen und Navigationsmuster, aber keinen direkten Einfluss auf den Wissensgewinn. Die Nutzer wiesen im Durchschnitt einen höheren Wissensgewinn durch Suchsitzungen zu Themen auf, die ihnen weniger bekannt waren.

- *Vorhersage des Wissenszuwachses der Benutzer in informativen Suchsitzungen.* Wir stellen ein überwachtes Modell vor, das den Wissensstand eines Benutzers und seinen Wissensgewinn durch die während der Suchvorgänge erfassten Funktionen vorhersagt. Wir verwenden einen umfassenden Satz von Feature-Sets und Feature-Selection-Strategien aus dem aktuellen Stand der Technik sowie unserer eigenen Forschung und vergleichen deren Leistung in unseren Modellen. Unsere Ergebnisse zeigen, dass es möglich ist, den Wissensstand und den Wissenszuwachs mit Hilfe von Merkmalen, die während der Suchvorgänge gesammelt wurden, vorherzusagen und zu klassifizieren. Unser Vorhersagemodell zeigt dabei eine verbesserte Leistung im Vergleich zu einer bestehenden Baseline für die Vorhersage des Wissensstandes.

- *Themenunabhängige Modellierung von Benutzerwissen in informativen Suchsitzungen.* Unsere Untersuchung zeigt, dass es möglich ist, überwachte Modelle zu erstellen, mit denen der Wissensgewinn und der Wissensstand eines Benutzers durch Benutzerinteraktionen während einer Suchsitzung vorhergesagt werden kann. Die Eigenschaften der Ressourcen, mit denen ein Benutzer interagiert, sind jedoch bei dieser Aufgabe weder ausreichend erforscht noch genutzt worden. Daher setzen wir unsere Forschung fort und stellen eine Reihe neuartiger ressourcenzentrierter Funktionen vor sowie ihre Fähigkeit, überwachte Modelle für die Vorhersage von Wissenszuwachs und Wissensstand der Nutzer in Web-Suchvorgängen deutlich zu verbessern. Unsere Beiträge sind wichtig, denn zuverlässige Trainingsdaten für solche Aufgaben sind spärlich und teuer zu beschaffen. Wir stellen verschiedene Strategien zur Featureauswahl vor, die darauf abzielen, eine spezifische Teilmenge von effektiven und verallgemeinerbaren Features zu selektieren.

Schlagworte: *Knowledge Base Augmentation, Suche als Lernen, Benutzermodellierung, Suchprotokollanalyse*

FOREWORD

The studies presented in this thesis have been published at various conferences or journals, as follows.

Chapter 3 is based on the works published in:

- Ran Yu, Ujwal Gadiraju, Besnik Fetahu, Oliver Lehmberg, Dominique Ritze and Stefan Dietze. KnowMore - Knowledge Base Augmentation with Structured Web Markup. In ***Semantic Web***: volume 10, issue 1, pages 159-180, 2019 (Journal Article). [YGF+19a]

In Chapter 4, we describe contributions included in:

- Ujwal Gadiraju, Ran Yu, Stefan Dietze. and Peter Holtz. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In ***CHIIR***: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 2-11, 2018. (Full Paper). [GYDH18]

Chapter 5 and 6 is built upon the works:

- Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting User Knowledge Gain in Informational Search Sessions. In ***SIGIR***: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 75-84, 2018 (Full paper). [YGH+18]

- Ran Yu, Rui Tang, Markus Rokicki, Ujwal Gadiraju, and Stefan Dietze. Topic-independent Modeling of User Knowledge in Informational Search Sessions. In ***WSDM***: *The 12th International Conference on Web Search and Data Mining*, 2019 (Full paper, under review). [YTR+19]

During the stages for my Ph.D. studies, I have also published a number of papers investigating different areas of Semantic Web technologies and Information Retrieval. Not all researched areas are touched in this thesis due to space limitation, but the complete list of publications follows:

- Ran Yu, Ujwal Gadiraju, Besnik Fetahu and Stefan Dietze. Adaptive Focused Crawling of Linked Data. In ***WISE***: *International Conference on Web Information Systems Engineering*, pages 554-569, 2015 (Full Paper). [YGFD15]

- Jakob Beetz, Ina Blmel, Stefan Dietze, Besnik Fetahu, Ujwal Gadiraju, Martin Hecher, Thomas Krijnen, Michelle Lindlar, Martin Tamke, Raoul Wessel and Ran Yu. Enrichment and Preservation of Architectural Knowledge. 3D Research Challenges in Cultural Heritage II - How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage, 2016 (Book Chapter). [BBD+16]

- Pracheta Sahoo, Ujwal Gadiraju, Ran Yu, Sriparna Saha and Stefan Dietze. Analysing Structured Scholarly Data Embedded in Web Pages. In ***SAVE-SD**: International Workshop on Semantic, Analytics, Visualization, 25th International Conference on World Wide Web*, 2016 (Workshop Paper). [SGY+16]

- Ran Yu, Besnik Fetahu, Ujwal Gadiraju and Stefan Dietze. A Survey on Challenges for Entity Retrieval in Web Markup Data. In ***ISWC**: 15th International Semantic Web Conference*, 2016 (Poster). [YFGD16]

- Ran Yu, Ujwal Gadiraju, Xiaofei Zhu, Besnik Fetahu and Stefan Dietze. Entity Summarization on Structured Web Markup. In ***ESWC**: 13th Extended Semantic Web Conference*, pages 69-73, 2016 (Poster). [YGZ+16]

- Stefan Dietze, Davide Taibi, Ran Yu, Phil Barker, and Mathieu d'Aquin. Analysing and Improving embedded Markup of Learning Resources on the Web. In ***WWW**: Proceedings of the 26th International Conference on World Wide Web Companion*, pages 283-292, 2017 (Full Paper). [DTY+17]

- Ran Yu, Ujwal Gadiraju, Besnik Fetahu, and Stefan Dietze. FuseM: Query-centric Data Fusion on Structured Web Markup. In ***ICDE**: 2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 179-182, 2017 (Poster). [YGFD17]

- Ujwal Gadiraju, Ran Yu, Stefan Dietze. and Peter Holtz. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In ***CHIIR**: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 2-11, 2018. (Full Paper). [GYDH18]

- Ran Yu, Ujwal Gadiraju and Stefan Dietze. Detecting, Understanding and Supporting Everyday Learning in Web Search. In ***WebSci**: Workshop on Learning & Education with Web Data (LILE), 10th ACM Conference on Web Science*, 2019 (Workshop Paper). [YGD18]

- Anett Hoppe, Peter Holtz, Yvonne Kammerer, Ran Yu, Stefan Dietze and Ralph Ewerth. Current Challenges for Studying Search as Learning

Processes. In ***WebSci**: Workshop on Learning & Education with Web Data (LILE), 10th ACM Conference on Web Science*, 2019 (Workshop Paper). [HHK+18]

- Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting User Knowledge Gain in Informational Search Sessions. In ***SIGIR**: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 75-84, 2018 (Full paper). [YGH+18]

- Ran Yu, Ujwal Gadiraju, Besnik Fetahu, Oliver Lehmberg, Dominique Ritze and Stefan Dietze. KnowMore - Knowledge Base Augmentation with Structured Web Markup. In ***Semantic Web**: Semantic Web*, volume 10, issue 1, pages 159-180, 2019 (Journal Article). [YGF+19a]

- Ran Yu, Rui Tang, Markus Rokicki, Ujwal Gadiraju, and Stefan Dietze. Topic-independent Modeling of User Knowledge in Informational Search Sessions. In ***WSDM**: The 12th International Conference on Web Search and Data Mining*, 2019 (Full paper, under review). [YTR+19]

- Masoud Davari, Ran Yu and Stefan Dietze. Understanding The Influence of Task Difficulty on User Fixation Behavior. In ***EARS**: The 2nd International Workshop on ExplainAble Recommendation and Search, 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019 (workshop paper). [DYD19]

- Ran Yu, Mathieu d'Aquin, Dragan Gasevic, Joachim Kimmerle, Eelco Herder and Ralph Ewerth. LILE2019: 8th International Workshop on Learning and Education with Web Data. In *Companion Publication of the 10th ACM Conference on Web Science*, 2019 (workshop summary). [YdG+19]

- Ran Yu, Ujwal Gadiraju, Besnik Fetahu, Oliver Lehmberg, Dominique Ritze, Stefan Dietze. In *The Semantic Web – The 18th International Semantic Web Conference (ISWC 2019) – Journal Track*, 2019 (extended abstract). [YGF+19b]

Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

This section gives the detailed motivation behind the two main lines of works introduced in this thesis, namely, *improving the accessibility of knowledge on the Web for machines* and *improving the accessibility of knowledge on the Web for humans*.

1.1.1 Improving the Accessibility of Knowledge on the Web for Machines

Technology is progressing rapidly, and it is changing the way of people accessing information in daily life. More and more applications have been developed with the goal to make it easier for human to access information. For instance, major search engines such as Google[1] and Bing[2] answer fact-checking queries by directly showing the answer in Search Engine Result Pages (SERPs), smart assistants such as Alexa and Siri answer users' requests in conversations. In most cases, machine-accessible knowledge is a prerequisite for building such smart applications. As of June 2019, there were over 1.3 billion websites on the Internet[3], which constitute a large collection of knowledge. However, most of the knowledge on the Web are unstructured and hard for machines to access directly. In order to aid smart applications, it is important to increase the knowledge accessibility for machines.

Knowledge bases (KBs) in this thesis refers to RDF datasets published based on a set of linked data principles introduced by Berners-Lee et al. [BLHL+01], which are also commonly referred to as *knowledge graphs*. KBs such as Freebase [BEP+08] or YAGO [SKW07] are in widespread use to aid a variety of applications and tasks such as Web search and smart assistant. While KBs capture large amounts of factual

[1] https://www.google.com/
[2] https://www.bing.com/
[3] According to the Netcraft Web Server Survey

knowledge, their coverage and completeness vary heavily across different types of domains. In particular, there is a large percentage of less popular (long-tail) entities and properties that are under-represented. For instance, at the time of our work presented in [YGF+19a], Freebase is missing statements for 63.8% (Wikidata for 60.9% and DBpedia for 49.8%) of all entities considering a selected set of properties used to describe books, such as language, publisher or number of pages (see Section 3.2). Here, gaps are in particular observable for less popular books or attributes, such as *translator* or *number of pages*.

Recent efforts in knowledge base augmentation (KBA) aim at exploiting data extracted from the Web to fill in missing statements. These approaches extract triples from Web documents [DGH+14a], or exploit semi-structured data from Web tables [RLB15, RLOB16]. After extracting values, data fusion techniques are used to identify the most suitable value (or fact) from a given set of observed values, for example, the correct director of a movie from a set of candidate facts extracted from the Web [DGH+14b]. To this end, data fusion techniques are fundamental when attempting to solve the KBA problem from observed Web data.

Although the extraction of structured data from Web documents is costly and error-prone, the recent emergence of embedded and structured Web markup has provided an unprecedented source of explicit entity-centric data, describing factual knowledge about entities contained in Web documents. Building on standards such as RDFa[4], Microdata[5] and Microformats[6], and driven by initiatives such as *schema.org*[7], a joint effort led by Google, Yahoo!, Bing and Yandex, markup data has become prevalent on the Web. An analysis from 2014 shows that 30% of all pages in a crawl of 2.01 billion HTML documents contain some form of embedded markup [MPB14], while this proportion has grown to 37.1% in 2018, considering a crawl of 2.5 billion documents[8]. This demonstrates a general upward trend of adoption, where the proportion of pages containing markup increased from 5.76% to 37.1% between 2010 and 2018.

Through its wide availability, markup lends itself as a diverse source of input data for KBA. In particular, when attempting to complement information about long-tail attributes and entities, the diversity and scale of markup provide opportunities for enriching existing knowledge bases and graphs [MRP16].

However, the specific characteristics of facts extracted from embedded markup pose particular challenges [YFGD16]. In contrast to traditional, highly connected RDF graphs, markup statements mostly consist of isolated nodes and small subgraphs, where entity descriptions often describe the same or highly related entities, yet are not linked through common identifiers or explicit links. For instance, in the WDC2013 corpus, 18,000 disconnected entity descriptions are retrieved when query-

[4] RDFa W3C recommendation: http://www.w3.org/TR/xhtml-rdfa-primer/
[5] http://www.w3.org/TR/microdata
[6] http://microformats.org
[7] https://schema.org/
[8] http://webdatacommons.org/structureddata/2018-12/stats/stats.html

ing the label of instances of type *schema:Product* for *'Iphone 6'*. Also, extracted markup statements are highly redundant and often limited to a small set of highly popular predicates, such as *schema:name*. Another challenge is data quality, as data extracted from markup contains a wide variety of errors, ranging from typos to the frequent misuse of vocabulary terms [MP15]. Our work in Chapter 3 aims at addressing the aforementioned challenges.

1.1.2 Improving the Accessibility of Knowledge on the Web for Humans

Web search is among the most frequent online activities and has become a ubiquitous task. Many search activities pertaining to the search for a particular piece of information expected to be available on the Web, are common and involve a particular learning intent, that is, the intent to acquire knowledge with respect to a certain topic. As is common search practice, a coherent search *session*, involving a particular search intent, usually involves several queries as well as one or more breaks in between (cf. [HGBS13]).

Whereas platforms dedicated to online learning, such as MOOC environments, are tailored towards improving the learning performance and experience of online users, contemporary search engines have to satisfy a range of use cases, which may or may not involve learning. In contrast to actual learning-oriented environments in the online or offline sphere, where certain knowledge about the learning intent, the user, the learning task as well as the suitable learning resource usually is available, such information is lacking in general online search settings. Consequently, heterogeneous features observable throughout a Web search session have to be utilised to derive insights about the learning intent, the user and the actual learning task. Furthermore, the findings should be integrated into the resource optimization, retrieval and ranking process in order to support user learning.

Recently, a range of research works have approached this problem, often summarised under the '*search as learning (SAL)*' umbrella and involving distinct disciplines such as information retrieval, human-computer interaction or machine learning.

Figure 1.1 summarises the key emerging research challenges which at the same time define the motivation of our works. *Detecting learning* in Web search, refers to the process of distinguishing learning-related activities from other, non-learning, activities in general Web search scenarios. *Understanding learning* refers to the challenges involved in inferring information about a user, such as her knowledge state, the learning task, such as its complexity, the learning process, or the involved resources from unstructured behavioral data observable throughout an online search session. Finally, *supporting learning* through retrieval, ranking and resource optimization refers to the actual consideration of inferred learning needs as part of the retrieval and ranking process, through adapting search interfaces to the user's learning intent, or provide more comprehensive learning resources.

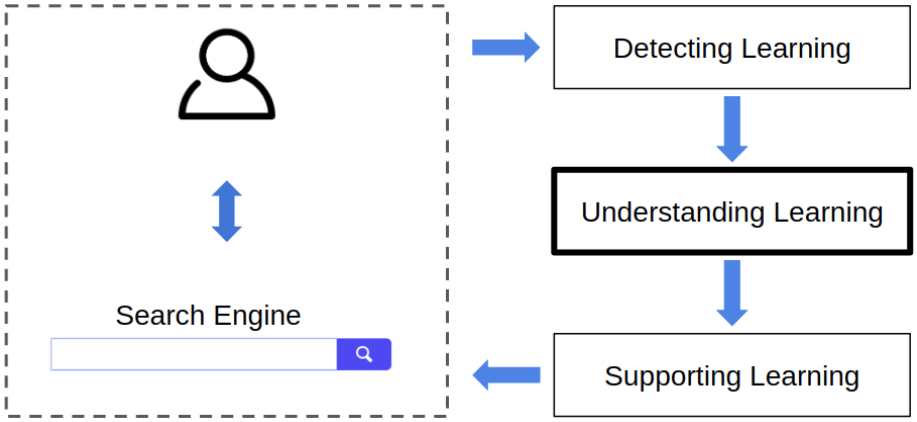


Figure 1.1 The detecting, understanding, supporting everday learning in Web search pipeline.

We worked on addressing all the 3 challenges as described above. For *detecting learning*, we adopted an established taxonomy from Broder [Bro02] and developed a classification approach to detect the intent of search sessions. Details about this work are described in [YGD18]. For supporting learning, our ongoing work explores the directions such as embedding interactive knowledge graph in SERPs and building collaborative search system.

This thesis focuses on *understanding human learning in Web search*. Recent research at the intersection of information retrieval and learning theory has recognized the importance of learning scopes and focused on observing the learning process during Web search. Eickhoff et al. investigated the correlation between several queries and search session-related metrics and learning progress [ETWD14]. Wu et al. predicted the difficulty of search tasks from query and session-related features [WKEA12]. Collins-Thompson et al. investigated the effectiveness of user interaction with respect to certain learning outcomes [CTRHS16]. In addition, Zhang et al. have shown that data obtained online during the search process provides valuable indicators about the domain knowledge of a user [ZCB11].

While prior works have focused on improving the learning experience and efficiency during search sessions, the measurement of a user's knowledge gain through the course of an informational search session has not yet been addressed. The importance of learning as an outcome of Web search has been recognized. Yet, there is a lack of understanding of the impact of Web search on a user's knowledge state. This is a vital cog in the wheel, if Web search engines that are currently optimized for relevance can be molded to serve learning outcomes. Our work in Chapter 4 has explored the correlation between Web search behavior and a user's knowledge state and knowledge gain (i.e., a user's learning performance).

Although we are able to extend the understanding of the relation between users' search behaviors and their knowledge gain through the course of an informational

search session, the automatic measurement of a user's knowledge gain that can be understood by search systems has not yet been addressed. This is in part due to the difficulty in accurately quantifying knowledge gain through the course of a search session. In order to re-molded the Web search engines to serve learning outcomes, the capability to predict knowledge gain will be a crucial step forward. Chapter 5 introduces our approach for the prediction of knowledge state as well as knowledge gain of a user using a range of behavioral signals captured during online search sessions and features extracted from user visited Web resources. The proposed features pertain to queries, sessions or behavioral traces, including mouse movements and navigational activities.

However, up to this point, our work has been constrained by limited and very specific feature sets. Insights into the generalizability of predictive models across topics are still shallow. This is particularly concerning in the light of our recent work in Chapter 4, which has found that the correlation between search behavior and search topic is stronger than the correlation between search behavior and the corresponding knowledge indicators (knowledge gain, knowledge state). Building on the observations in our previous works, in Chapter 6, we explore further by introducing a novel set of Web resource-centric features and investigate their impact on the knowledge gain/state prediction task. We introduce various feature selection strategies geared towards selecting a limited subset of effective and generalizable features by considering feature correlation with knowledge gain/state, topic-dependency of feature performance and feature redundancy.

1.2 Contributions of this Thesis

In this thesis, we address the challenges described in Section 1.1.1 through *knowledge base augmentation with structured Web markup* (contribution **(I)**) and the challenges described in 1.1.2 through a series of works in the scope of *understanding human learning in Web search* (contribution **(II)**, **(III)** and **(IV)**).

(I) ***Knowledge Base Augmentation with Structured Web Markup:*** In Chapter 3, we introduce *KnowMore*, an approach based on data fusion techniques which exploits markup crawled from the Web as diverse source of data to aid KBA. Our approach consists of a two-fold process, where first, candidate facts for augmentation of a particular KB entity are retrieved through a combination of blocking and entity matching techniques. In a second step, correct and novel facts are selected through a supervised classification approach and an original set of features. We apply our approach to the WDC2015 dataset and demonstrate superior performance compared to state-of-the-art data fusion baselines. We also demonstrate the capability for augmenting three large-scale knowledge bases, namely Wikidata[9],

[9] https://www.wikidata.org/

Freebase and DBpedia[10] through markup data based on our data fusion approach. The main contributions of our work are threefold:

- **Pipeline for data fusion on Web markup.** We propose a pipeline for data fusion (Section 3.2.3) that is tailored to the specific challenges arising from the characteristics of Web markup (Section 3.2.1). In particular, given the dynamics and scale of markup data, our approach performs the task of query-centric data fusion, which provides an efficient means to fuse only specific parts of a given markup corpus, obtained through a preliminary blocking step. Relevance, diversity, and correctness of facts is addressed through a combination of entity matching, and data fusion techniques. To the best of our knowledge, this is the first approach addressing the task of data fusion on Web markup data.
- **Model & feature set.** We propose a novel data fusion approach consisting of a supervised classification model (Section 3.4), utilising an original set of features geared towards validating correctness and relevance of markup facts. Experimental results demonstrate high precision (avg. 91.7%) and recall (avg. 88.2%) of our model, outperforming the state-of-art baselines.
- **Knowledge base augmentation from markup data.** As part of our experimental evaluation (Section 3.6), we demonstrate the use of fused markup data for augmenting three well-established knowledge bases. Our results underline the suitability of markup data for supporting KBA tasks, where *KnowMore* is able to reach 100% coverage gain (Section 3.5.2) for selected properties and types, for instance, book descriptions in Freebase and Wikidata. On average, KnowMore has a coverage gain of 36.49% in Wikidata, 39.42% in Freebase and 34.75% in DBpedia. We also investigate the particular potential for augmenting tail entities and properties in Section 3.8.1.

(II) ***Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web:*** In Chapter 4, we describe novel insights on the nature of knowledge gain in informational search sessions on the Web, and the corresponding behavior of users. By combining qualitative and quantitative analysis, we seek to answer the following research questions.

- **RQ1: How does a user's knowledge evolve through the course of an informational search session on the Web?** To further the current understanding of the impact of informational search on a user's knowledge, we recruited 500 distinct users from a crowdsourcing platform and orchestrated search sessions spanning 10 different information needs. By employing scientifically formulated *knowledge tests* to calibrate a user's knowledge before a search session, and assess it after the session, we were able to quantify knowledge gain. We found that nearly 70% of the users exhibited a knowledge gain

[10] http://dbpedia.org

at the end of a search session corresponding to an information need, with an overall average knowledge gain of almost 20%.

- **RQ2: How does the information need in a search session influence a user's knowledge gain?** We explored the impact of information need on the knowledge gain of users. Our findings revealed that the information need does not directly affect the knowledge gain of users. However, we found a strong negative linear relationship between the knowledge gain of users in an informational search session and their topic familiarity. This suggests that users exhibited a higher knowledge gain in search sessions corresponding to information needs that they were less familiar with.
- **RQ3: What is the impact of information need on the search behavior of users in a search session?** We analyzed the search behavior of users and found a significant effect of the information need on the number of queries entered by users, the number of unique terms in their queries, the number of webpages that users navigated to, and the distinct pay-level domains accessed. Information need also had a significant effect on the amount of time users actively spent on the search results page. We also found that on average the last queries entered by users were significantly longer than the first queries across all information needs, suggesting an impact of the information consumed through the course of a search session.

(III) ***Predicting User Knowledge Gain in Informational Search Sessions:*** In Chapter 5, we introduce a supervised model to predict a user's knowledge state and knowledge gain from features captured during the search sessions. Through our work in this chapter, we make the following contributions to the current body of literature:

- **Novel feature sets.** A novel set of user behavioral features extracted from different dimensions of a search session, namely features related to the session, queries, SERP, browsing behavior and mouse movements.
- **Knowledge state/gain prediction models.** Models for predicting the user's knowledge gain and state during real-world informational search sessions. The experimental results underline that a user's knowledge gain and knowledge state can be modeled based on a user's online interactions observable throughout the search process.
- **Feature analysis.** An analysis of the effect of user interactions (ranging from the queries entered to their browsing behavior) on their knowledge state and knowledge gain.

(IV) ***Topic-independent Modeling of User Knowledge in Informational Search Sessions:*** In Chapter 6, we introduce a novel set of Web resource-centric features and investigate their impact on the knowledge gain/state prediction task. We introduce various feature selection strategies geared towards selecting a limited subset of

effective and generalizable features by considering feature correlation with knowledge gain/state, topic-dependency of feature performance and feature redundancy. In summary, our contributions include the following:

- **Novel feature sets.** We introduce and experimentally evaluate novel Web resource feature sets (109 features in total) for the task of knowledge state (KS) and knowledge gain (KG) prediction, which extends state of the art models.
- **Feature analysis.** We conduct comprehensive feature analysis assessing both generalisability of features across search topics as well as their overall effectiveness in the aforementioned prediction tasks. Findings from this analysis can inform future work for user modeling in search sessions in various ways. Moreover, our analysis can be leveraged to build computationally efficient models through a limited set of effective features.
- **Feature selection approach.** In order to cope with the wide variety and large number of features in the presence of very sparse training data, we introduce a novel approach for feature selection which combines feature correlation with target variables (KG/KS) as well as the topic-dependency of feature performance. By doing so, we identify the best performing features in cross-topic prediction settings and facilitate generalisable models.
- **Improved prediction models.** We evaluate our features and feature selection approach by building supervised classifiers which outperform state-of-the-art baselines for the knowledge gain/state prediction on *unseen* topics. On average, our improved models outperform the previous state-of-the-art baseline [YGH+18] by 20.6%, 39.9%, and 16% (average F1 score) in the tasks of knowledge gain, knowledge state, and post-knowledge state prediction, respectively.

2 Background

This chapter introduces the background necessary to understand the work carried out in accordance with this thesis. We first introduce the notion of *knowledge graph* and some of its important applications, then continue on *structured Web markup*. In the last part, we discuss *human learning in Web search* and introduce the recently emerged research fields of *search as learning*.

2.1 Knowledge Base

Knowledge bases (KBs) are datasets containing structured data about real-world entities. In this thesis, with the term "knowledge bases" we explicitly refer to RDF datasets that contain machine-readable knowledge. KBs are also commonly referred to as "knowledge graphs" (KGs). Although the term "knowledge base" and "knowledge graph" have often been used interchangeably, they are not necessarily synonymous according to existing definitions of both terms. Ehrlinger and Wöß [EW16] reviewed the definitions and mentions of "knowledge graph" in previous literatures, and proposed the definition: "*A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge*". They argue that the major difference between a KB and a KG is that a KB does not necessarily applies a reasoning engine to generate new knowledge. Hence, authors suggest that it is suitable to replace the term knowledge graph with knowledge base, but not vice versa.

2.1.1 RDF Data Model

The Resource Description Framework (RDF)[1] is the main standard for knowledge representation that enables the publishing of KBs. RDF models represent entities (i.e.

[1] W3C. Rdf schema 1.1. https://www.w3.org/TR/rdf-schema/

resources) by Uniform Resource Identifiers (URIs) and describe the resources using statements in the form of $\{subject, predicate, object\}$ triples. The *subject* denotes the resource; the *predicate*, which is usually referred to as *property*, is vocabulary of predefined schemas, denotes traits or aspects of the resource, and expresses a relationship between the *subject* and the *object*; *object* denotes the property value of the resource, which could be a URI of another entity or literals. This linking structure forms directed and labeled graphs, where the edges represent the semantic links between nodes (i.e. resources).

RDF schema (RDFs) is an extension of the basic construct provided by the RDF data model, it enables the construction of *classes*. Resources can be assigned to classes through the *rdf:type* property. Classes can be organized into hierarchical structure using attribute *rdfs:subClassOf*, such as in Listing 2.1, where *dbo:Film* is defined as a *class* and subclass of *dbo:Work*. By traversing the class hierarchy it can be inferred that *dbo:Film* is also a subclass of *owl:Thing*.

Listing 2.1 RDFs class definition example

```
dbo:Film rdf:type rdfs:class .
dbo:Film rdfs:subClassOf dbo:Work .
dbo:Work rdfs:subClassOf owl:Thing .
```

Listing 2.2 shows a snippet of DBpedia, which contains 5 facts of the movie *Forrest Gump* (i.e. *dbr:Forrest_Gump*), and 4 facts about its director *Robert Zemeckis* (i.e. *dbr:Robert_Zemeckis*). The 5 triples correspond to "*Forrest Gump*" provide information about its name, runtime, director, writer and entity type using DBpedia ontology (e.g. *dbo:director*) and rdf property (e.g. *rdf:type*). The predicate *dbo:director* creates a semantic link between the two entities.

Listing 2.2 Snippet of DBpedia

```
dbr:Forrest_Gump dbo:label "Forrest Gump" .
dbr:Forrest_Gump dbo:runtime "142.0"^^ns26:minute .
dbr:Forrest_Gump dbo:director dbr:Robert_Zemeckis .
dbr:Forrest_Gump dbo:writer dbr:Eric_Roth .
dbr:Forrest_Gump rdf:type dbo:Film .
dbr:Robert_Zemeckis dbo:label "Robert Zemeckis" .
dbr:Robert_Zemeckis dbo:birthDate "1952-05-14"^^xsd:date .
dbr:Robert_Zemeckis dbo:birthPlace dbr:Chicago .
dbr:Robert_Zemeckis rdf:type dbo:Person .
```

2.1.2 Current State and Applications of Knowledge Bases

With the power of providing machine-readable knowledge by integrating large amount of data from various data sources so that it can be used in a meaningful and more intelligent way, knowledge bases have becoming more and more prevalent for integrating and preserving data.

According to the report released by Linked Open Data Cloud[2], as of March 2019, 1239 datasets have been published according to the linked data principles. In the meantime, companies such as Facebook, Microsoft, Google operate their own large scale knowledge bases as part of their infrastructure for commercial use. Among the existing knowledge bases, DBpedia [ABK+07], Wikidata [Vra12], freebase [BEP+08] and Yago [SKW07] have been recognized as the largest public accessible cross-domain knowledge bases with number of statements ranging from 63.2 million to billions and have been used extensively for research purposes. Färber et al. [FEMR15] provided a detailed analysis and comparison of these 4 knowledge bases. Despite of the fast growing scale, many issues exist in current knowledge bases. One of the challenges attracts most attention is the incompleteness of the KBs. In particular, there is a large percentage of less popular (long-tail) entities and properties that are under-represented. Improving the completeness of the KBs is the main focus of our work introduced in Chapter 3.

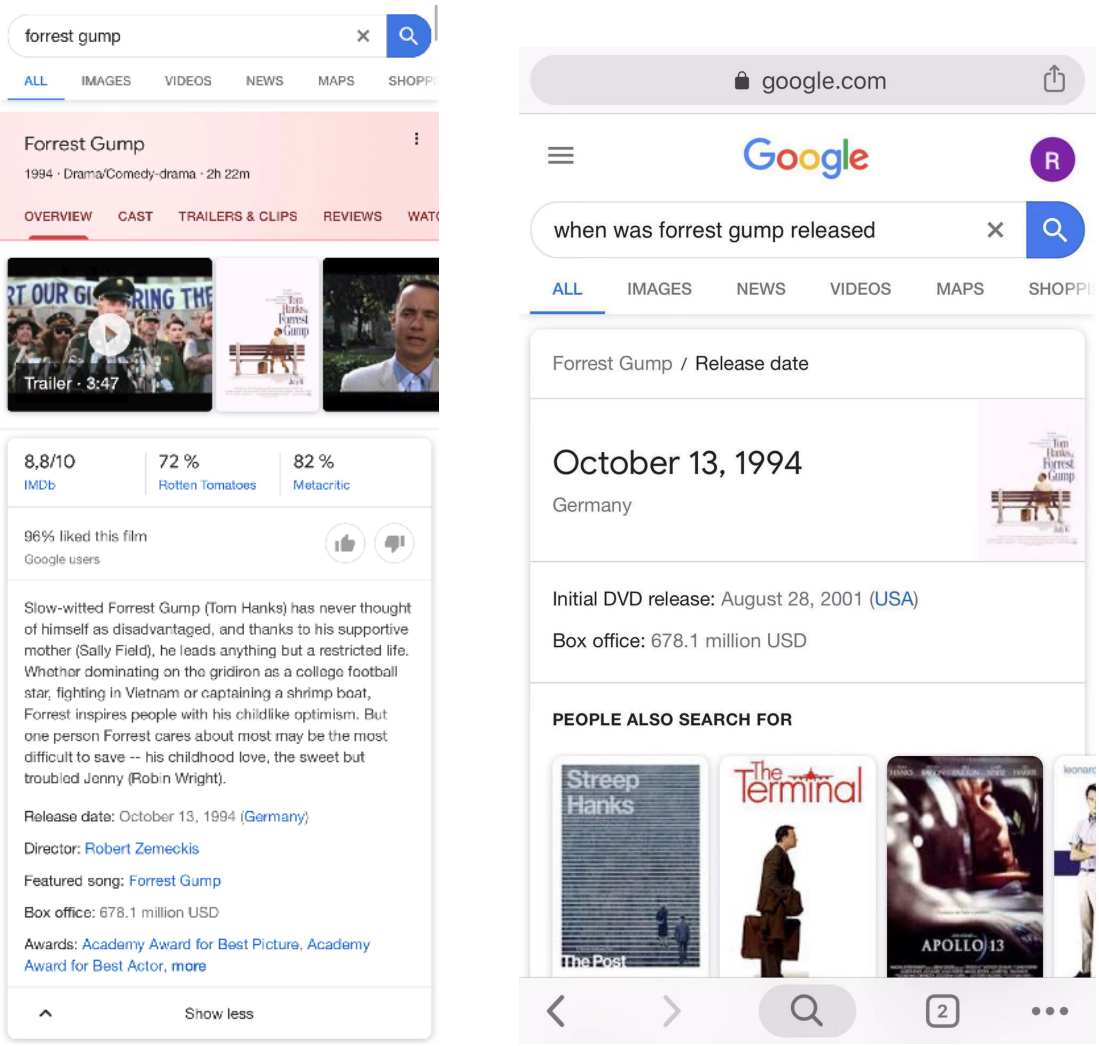
KBs have been used for improving the performance of many prevalent applications such as Web search, question answering, smart assistant as well as product recommendations. Figure 2.1a is a screenshot taken from Google search engine result page (SERP) of query "Forrest Gump", for which *Google Knowledge Graph* [Sin12] has been used to enrich the result and directly show the key information of the query entity on SERP. In Figure 2.1b, Google search engine shows the answer to the question in user query and other information about the entity directly on SERP in a structured way. Companies providing smart assistant services (e.g. *Amazon Alexa*, *Siri* or *Google Assistant*), are also paying more and more attention on using KBs to improve the performance of their products. Furthermore, e-commerce platforms use KBs to improve the visibility of product information, for instance, *amazon* is building the *Amazon Product Graph* as an authoritative KB for products, with the goal of answering any question about products and related knowledge.

All the applications utilizing structured machine-readable knowledge that users are constantly interacting with on a daily basis, making it a crucial task to improve the completeness and quality of KBs.

2.2 Structured Web Markup

As discussed in section 2.1, more and more data has been integrated into KBs, providing large scale machine-readable knowledge that can support various applications. However, KBs still only constitute a small fraction of the Web. Internet user constantly interact with various types of Web resources other than KBs. Hence it is important for systems such as search engine to be able to better understand the content of other type of Web resources. To this end, standards such as *Microdata*[3],

[2] https://lod-cloud.net
[3] https://www.w3.org/TR/microdata/



(a) Entity-centric query **(b)** Question as query

Figure 2.1 Example of using KG for enriching search result (screenshots taken on smart phone).

Microformats[4] and *RDFa*[5] have been established to enable the embedding of structured entity information in webpages. The embedded information about entities in webpages is expressed in the same way as in knowledge bases – each fact can be extracted to a triple which consists of a subject, a predicate and an object. Figure 2.2 shows an exmaple of using *microdata* attributes to add entity descriptions into a IMDB page[6], where explicit facts about the actor *Tom Hanks* are embedded in the way that is easy to be parsed by machines.

[4] http://microformats.org/
[5] http://rdfa.info/
[6] https://www.imdb.com/title/tt0109830/

Figure 2.2 Example of using structured Web markup to embed entity information.

Driven by search engine providers such as Google, Yahoo!, Bing and Yandex, structured Web markup has reached significant adoption, where more than 37% of all Web pages in Common Crawl[7], the largest web corpus available to the public, already provide some form of markup as of November 2018[8]. As such, markup constitutes a source of entity-centric data on the Web at an unprecedented scale. As a large collection of structured entity-centric data, it has the potential of enriching KBs. In Chapter 3, we introduce our approach of using Web markup data for knowledge base augmentation. These semantic annotations have been used by search engines and online retailer etc. to improve their search performance and to have richer display of results within their applications. The structured knowledge embedded in webpages also make the access of knowledge easier for humans through various applications.

2.2.1 Markup Standards

In this section, we introduce the three most prevalent markup standards used by website administrators and accepted by the major search engines, namely, *Microformats*, *RDFa* and *Microdata* [Meu17].

Microformats are a set of open data formats built upon existing and widely adopted standards, which use existing HTML/XHTML tags (*class*, *rel*, *rev*) to embed

[7] https://commoncrawl.org/
[8] http://webdatacommons.org/structureddata/2018-12/stats/stats.html

entity information into webpages. It supports a list of formats (i.e. classes) and defines a set of properties for each class. Microformats annotations cannot be combined with other vocabularies.

Some of the most popular classes and example properties for describing instances of each class are listed in Table 2.1. The full list of Microformats can be found online[9]. Listing 2.3 shows a HTML snippet that uses Microformats to embed metadata of an entity of class *vcard*. In the example, the photo, name, url and address information of the person *Robert Zemeckis* are annotated with semantics. As shown in the example, Microformats do not differentiate between class and property, which could bring ambiguity to software systems when parsing data.

Table 2.1 Examples of Microformats classes and properties.

Class	Domain	Properties
geo	locations	*latitude, longitude*
vcard	people, contacts and organization	*url*, *fn* (name), *photo*
vcalendar,vevent	calendars and events	*location*, *url*, *location*
hlisting	listings for products or services	*description*, *price*, *version*
hrecipe	cooking and baking recipes	*fn* (name), *ingredient*, *duration*

Listing 2.3 Microformats example

```
<p class="vcard">
  <img class="photo" src="http://example.org/photo.png" alt="" />
  <a class="url fn" href="http://example.com/">Robert Zemeckis</a>
  <span class="country-name">USA</span>
</p>
```

RDFa (Resource Description Framework in Attributes) is a set of attribute-level extensions to HTML, XHTML and various XML-based document types for embedding metadata in Web documents. The set of extensions includes *vocab*, *prefix*, *resource*, *property* and *typeof*, which support the modeling of almost all RDF expressions. RDFa also supports the use of other existing vocabularies through additional attributes such as *about*, *src*, *href*, *content*, *datatype*, *inlist*, *rel* and *rev*.

Listing 2.4 RDFa example

```
<div vocab="http://schema.org/" typeof="Person">
  <div property="name">Robert Zemeckis</div>
  <div property="birthDate">1952-05-14</div>
  <div property="birthPlace">Chicago</div>
</div>
```

[9] http://microformats.org/wiki/Main_Page

Listing 2.4 shows an example snippet of using RDFa attributes and *schema.org* vocabularies to embed entity information. Unlike Microformats, RDFa standard differentiate between class and property, which brings more semantics into Web annotations. The entity is described using *schema.org*[10] vocabularies, which were initiated by the major search engines such as Google, Microsoft, Yahoo!, and Yandex and became the most dominant vocabularies for structured Web markup [MPB14].

Microdata is an open-community HTML specification for embedding structured metadata into Web documents. It defines a list of global attributes that extend standard HTML attributes for describing items and property-value pair, including *itemscope*, *itemtype*, *itemid*, *itemprop*, *itemref* and *datetime*. Microdata does not provide vocabularies for describing the semantics of instances directly, but supports the use of existing or custom vocabularies.

Listing 2.5 Microdata example

```
<section itemscope itemtype="http://schema.org/Person">
  <span itemprop="name">Robert Zemeckis</span>
  <span itemprop="birthDate">1952-05-04</span>
  <span itemprop="birthPlace">Chicago</span>
</section>
```

Listing 2.5 shows an example of using Microdata attributes to add metadata of an entity of type *Person*, where explicit information about *Robert Zemeckis* is embedded into the Web document using attributes *itemtype* and *itemprop*.

2.2.2 Characteristics of Web Markup Data

Since year 2012 [MB12], the *Web Data Commons* (WDC) project[11] extracts embedded structured data from webpages in the Common Crawl and provides the extracted data in *N-Quads*[12] format for public download yearly. According to the reports from WDC, the percentage of webpages (counted by number of URLs) having embedded markup has grown from 12.3% to 37.1% between 2012 and 2018.

With the potential of providing very large scale machine-readable knowledge for enriching KBs, the quality issues of the markup data have downgraded its usability. Meusel et al. [MPB14, MP15, MRP16] investigated the common errors exist in WDC datasets, and found that 1.23% of 398,542 pay-level domains (PLDs) deploy at least one wrong namespace meant to be *http://schema.org*, 6.07% of PLDs make use of undefined *schema.org* types, 3.92% of PLDs use at least one undefined *schema.org* property, and 56.58% of PLDs use object properties (i.e. the property value should be an URI refers another entity) defined by *schema.org* as datatype property (i.e. with a literal property value) at least once.

[10] https://schema.org/
[11] http://webdatacommons.org/
[12] https://www.w3.org/TR/n-quads/

Apart from the common errors discussed by Meusel et al. [MP15], more complex characteristics of facts extracted from embedded markup pose particular challenges. In our previous work [DTY+17], we investigated the adoption of *Learning Resource Metadata Initiative* (LRMI) (i.e. vocabularies for annotating learning resources through *schema.org* terms) on the Web. Based on the analysis result of LRMI markup in WDC datasets from three consecutive years (2013-2015), we found that LRMI vocabularies have been semantically misused by many Web documents. For instance, the property *schema:typicalAgeRange* is often used by websites providing adult content. In [YFGD16], we presented a preliminary analysis of challenges in using Web markup data for entity retrieval. The result shows that statements extracted from Web markup mostly consist of isolated nodes and small subgraphs, where entity descriptions often describe the same or highly related entities, yet are not linked through common identifiers or explicit links. Also, extracted markup statements are highly redundant and often limited to a small set of highly popular properties, such as *schema:name*.

In summary, in order to mining machine-readable knowledge from structured Web markup, we first need to overcome the challenges that the data is noisy, redundant and the nodes are isolated. More details about the challenges in mining structured knowledge from Web markup and our approach aims at overcoming these challenges are presented in Chapter 3.

2.3 Human Learning in Web Search

A Web search system for Web resources, also commonly referred to as *search engine* or *information retrieval system*, is a software system that aims at satisfying user information need through retrieving resources from a large Web corpus. The information need of a user is expressed through search queries, the corresponding search results are generally returned to the user as ranked lists with supportive information on dynamically generated webpages, often referred to as search engine result pages (SERPs).

The first search engine *Archie* was created on 1990 as a tool for finding files in FTP archives, throughout the years, search engine has developed into one of the most powerful and commonly used tools for accessing information on the Web. Search engines have gone through massive optimizations from different aspects such as personalized ranking, query suggestion, multimodal search and semantic search (i.e. search system that understands the searcher's intent and the contextual meaning of terms). Many closely related techniques such as user modeling and nature language processing are also advanced because of their adoption in search engines.

Although current search engines can fulfill users' information needs effectively from a relevance-based perspective, there are still challenges and opportunities in satisfying more complexed information needs. Recently, more and more attention

has been spent on understanding the relation between the behavior of users, their information need, Web resources and users' information gain. Research fields such as interactive information retrieval (IIR), human-computer information retrieval (HCIR) and search as learning (SAL) have became active. Our work presented in Chapter 4, 5 and 6 of this thesis focus on the SAL scenario, specifically, understanding the learning process of search engine users.

Search as Learning (SAL). Whereas platforms dedicated to online learning, such as MOOC environments, are tailored towards improving the learning performance and experience of their users. Unlike learning platforms, contemporary search engines have to satisfy a range of use cases, which may or may not involve learning. In contrast to actual learning-oriented environments in the online or offline sphere, where certain knowledge about the learning intent, the user as well as the learning task usually is available, such information is lacking in general Web search settings. Consequently, heterogeneous features observable throughout a Web search session have to be utilized to derive insights about the learning intent, the user and the actual learning progress.

Recently, a range of research works have approached this problem, often summarized under the '*search as learning (SAL)*' umbrella. Given the extensive usage of search engine in everyday learning, there has been a growing recognition of the importance of studying and designing search systems to foster discovery and enhance the learning experience during the search process outside of formal educational settings [CTHH17]. SAL related topics have been studied by researchers from distinct disciplines such as learning analytics, education, Web mining, Web science, psychology and the social sciences. In the following, we discuss the SAL related theories and technologies that are introduced in existing literatures.

2.3.1 Learning Types

Many different learning taxonomies have been defined in the scope of psychology, however, there is no widely recognized definition of learning or taxonomy of learning types in the context of SAL.

Previous SAL related works either consider "learning" as an intuitive concept that does not need to be defined (e.g. [CGL+13, Aro15]), or focus only on very specific scenario of learning activities (e.g. [ETWD14, SCT18]). Eickhoff et al. [ETWD14] considered two types of knowledge acquisition intent – procedural knowledge and declarative knowledge. They extract corresponding search engine log by selecting sessions with query terms such as "how to" for procedural knowledge and "what is" for declarative knowlege. Syed et al. [SCT17, SCT18] focus specifically on the vocabulary learning scenario and carry out analysis based on data collected through lab studies. Hagen et al. [HPV+16] investigated the relation between the writing behavior and the exploratory search pattern of writers.

Although previous works did not use a common definition of learning, most of

the investigated activities are associated with *intentional learning*, which is generally defined as *"learning that is motivated by intentions and is goal directed,in contrast to latent or incidental learning"* [BS+89, Blu12]. Another take on learning in Web search by Rose et al. [RL04] posits that a search activity is intentional learning related, when *"a user's goal is to learn something by reading or viewing Web pages"*.

For the categorization of search queries, a widely used taxonomy from Broder [Bro02] distinguishes between *transactional*, *navigational* and *informational* queries. Specifically, queries with *transactional* intent usually aim at conducting a specific online transaction, such as, purchasing a ticket, *navigational* queries merely are aimed at leading the user to a dedicated website. In contrast, *informational* queries imply the intent of a user to acquire some knowledge assumed to be present on one or more webpages, which aligns with the definition of intentional learning as mentioned earlier.

A coherent search *session*, involving a particular search intent, usually involves several queries as well as one or more breaks in between (cf. [HGBS13]). Recent studies have shown that information seeking tasks have grown more sophisticated [JK08] and often require one or more queries across multiple search sessions [KBW+11, AWDB12]. Hence the intent behind search activities should be regarded at session level rather than a single query. On this basis, we manually inspected a real-world query log, which consists of 913 coherent search sessions, and found that 49.7% of them were informational search sessions with specific learning intent. Due to the learning intention behind and the prevalence of informational search sessions, we focus on the investigation of such sessions in the context of improving the accessibility of knowledge on the Web for humans.

2.3.2 Learning Process

The Anderson and Krathwohl's taxonomy [AKA+01] has been adopted by several SAL related works for designing lab studies (e.g. [KAEW15]). It defines 6 levels of the cognitive process, namely, remembering, understanding, applying, analyzing, evaluating and creating. In a more recent survey, Vakkari [Vak16] proposed a systematization of the learning process in search, which consists of four stages: search formulation, selecting sources, interacting with sources and presenting. As no explicit indicators have been found so far that can be used to align learning activities with above defined learning stages, many recent works (e.g. [ZCB11, Arg14, YGH+18, SCT18]) considered learning as a continuously process rather than discrete stages.

2.3.3 Challenges in SAL

According to prior literatures [CTHH17, Vak16, YGD18] and our own experience, we summarize the key building blocks of SAL as in Figure 1.1 and the research challenges correspond to each block as follows:

- **Detecting learning** refers to the process of distinguishing intentional learning-related activities from other non-intentional learning activities in general Web search scenarios. As search engines have to satisfy a range of use cases, which may or may not involve learning intent, making it a prerequisite to identify the learning related search activities before any learning-oriented optimizations can be applied. Furthermore, it is important that the developed approaches are generalizable and can be applied in real world search engines in real-time.

- **Understanding learning** refers to tasks involved in inferring information about a user (e.g. her knowledge state), the learning task (e.g. its complexity), the learning progress and the influence of involved resources from unstructured behavioral data observable throughout an online search session. The understanding of the SAL process can help identifying the optimization directions of search engines and provide supportive information for optimization approaches.

- **Supporting learning** is the end goal of SAL related tasks and many directions can be explored for realizing it. The 3 major possibilities are: 1) through retrieval and ranking by considering the inferred learning needs as part of the retrieval and ranking process, 2) through adapting search interfaces to improve learning efficiency, and 3) provide more comprehensive learning resources.

In our work we consider the *informational search* sessions as intentional learning sessions, and proposed a session classification approach[YGD18] for addressing the challenge of detecting learning. Chpater 4, 5 and 6 in this thesis introduce our effort on the challenge of understanding learning in Web search.

3

KnowMore – Knowledge Base Augmentation with Structured Web Markup

As discussed in Chapter 2, Knowledge Bases as machine-readable knowledge source on the Web, have been used widely in many applications such as Web search and smart assistant. However, their coverage and completeness vary heavily across different types or domains. In particular, there is a large percentage of less popular (long-tail) entities and properties that are under-represented. In this chapter, we focus on improving the knowledge accessibility of machines on the Web through knowledge base augmentation.

The recent emergence of embedded and structured Web markup has provided an unprecedented source of explicit entity-centric data, describing factual knowledge about entities contained in Web documents. Through its wide availability, markup lends itself as a diverse source of input data for KBA. In particular when attempting to complement information about long-tail attributes and entities, the diversity and scale of markup provide opportunities for enriching existing knowledge bases and graphs [MRP16].

However, the specific characteristics of facts extracted from embedded markup pose particular challenges [YFGD16]. In contrast to traditional, highly connected RDF graphs, markup statements mostly consist of isolated nodes and small sub-graphs, where entity descriptions often describe the same or highly related entities, yet are not linked through common identifiers or explicit links. Also, extracted markup statements are highly redundant and often limited to a small set of highly popular predicates, such as *schema:name*. Another challenge is data quality, as data extracted from markup contains a wide variety of errors, ranging from typos to the frequent misuse of vocabulary terms [MP15].

In this chapter, we introduce *KnowMore*, an approach based on data fusion techniques which exploits markup crawled from the Web as diverse source of data to aid KBA. Our approach consists of a two-fold process, where first, candidate facts

for augmentation of a particular KB entity are retrieved through a combination of blocking and entity matching techniques. In a second step, correct and novel facts are selected through a supervised classification approach and an original set of features. We apply our approach to the WDC2015 dataset and demonstrate superior performance compared to state-of-the-art data fusion baselines. We also demonstrate the capability for augmenting three large-scale knowledge bases, namely Wikidata[1], Freebase and DBpedia[2] through markup data based on our data fusion approach. The main contributions of our work are threefold:

- **Pipeline for data fusion on Web markup.** We propose a pipeline for data fusion (Section 3.2.3) that is tailored to the specific challenges arising from the characteristics of Web markup (Section 3.2.1). In particular, given the dynamics and scale of markup data, our approach performs the task of query-centric data fusion, which provides an efficient means to fuse only specific parts of a given markup corpus, obtained through a preliminary blocking step. Relevance, diversity, and correctness of facts is addressed through a combination of entity matching, and data fusion techniques. To the best of our knowledge, this is the first approach addressing the task of data fusion on Web markup data.

- **Model & feature set.** We propose a novel data fusion approach consisting of a supervised classification model (Section 3.4), utilising an original set of features geared towards validating correctness and relevance of markup facts. Experimental results demonstrate high precision (avg. 91.7%) and recall (avg. 88.2%) of our model, outperforming the state-of-art baselines.

- **Knowledge base augmentation from markup data.** As part of our experimental evaluation (Section 3.6), we demonstrate the use of fused markup data for augmenting three well-established knowledge bases. Our results underline the suitability of markup data for supporting KBA tasks, where *KnowMore* is able to reach 100% coverage gain (Section 3.5.2) for selected properties and types, for instance, book descriptions in Freebase and Wikidata. On average, KnowMore has a coverage gain of 36.49% in Wikidata, 39.42% in Freebase and 34.75% in DBpedia. We also investigate the particular potential for augmenting tail entities and properties in Section 3.8.1.

3.1 Related Work

In this section we review related literature. We focus on two main lines of work, namely *knowledge-base augmentation* and *data fusion* as the most closely related fields to our work.

[1] https://www.wikidata.org/
[2] http://dbpedia.org

3.1.1 Knowledge-base Augmentation (KBA)

The main goal of KBA is to discover facts pertaining to entities and augmenting Knowledge Bases (KB) with these facts [WT10, JG11].

Some previous works have proposed to augment KBs through inference on existing knowledge, i.e. KB statements. Such works typically focus on predicting the type [LABT11] of an entity or finding new relations based on existing data [BL12, BL13, SCMN13]. Other prior works are more closely relevant to our problem setup; in that they focus on predicting relations with external data. Notable works propose the use of Wikipedia as a text corpus annotated with entities, search for patterns based on existing KB relations, and further apply the patterns to find additional relations for DBpedia [AGL13] or Freebase [MBSJ09]. News corpora have also been used for augmenting DBpedia through similar approaches [GHB+13]. Paulheim et al. [PP13] proposed to identify common patterns of instances in Wikipedia list pages and apply the patterns to add relations to the remaining entities in the list. Dong et al. proposed '*Knowledge Vault*' [DGH+14a], a framework for extracting triples from webpages, aimed at constructing a KB from Web data. Dutta et al. [DMS15] focus on the mapping of relational phrases such as facts extracted by '*Nell*' and '*Reverb*' to KB properties. Furthermore, they group the same semantic relationships represented by different surface forms together through Markov clustering. Recent works by Ritze et al. use relational HTML tables available on the Web to fill missing values in DBpedia [RLB15, RLOB16]. The authors propose to first match the tables to the DBpedia entities, and then compare several data fusion strategies such as voting and the Knowledge-Base Trust (KBT) score to identify valid facts.

Ristoski et al. proposed an approach to enrich product ads with data extracted from the Web Data Commons [RM16]. The approach extracts attribute-value pairs from plain text and matches them to database entities with supervised classification models. The notable methods described in previous works are tailored to specific data sources, which have different characteristics compared to markup data. Hence, merely adopting the existing methods to cater for markup data is not sufficient. However, we have revised and adopted some of the features in our proposed approach.

Other works suggest using the whole Web as a potential data corpus through search engines [KM12, WGM+14]. QA-based approaches are often designed to facilitate the filling of values of a specific set of properties, and rely on manually created templates. This limits their application to a constrained sets of properties.

Existing works typically assume that, there is only one true value for a property when resolving conflicts. In contrast, we aim for higher recall by allowing multiple (non-redundant) correct values, catering for the fact that multiple-cardinality properties are wide-spread. Another limitation of existing KBA works is that the novelty of the discovered facts is ignored; there is an overlap between the result and the facts existing in a KB. On the contrary, our approach aims at providing correct and novel results that are of immediate value to the KB.

3.1.2 Data Fusion

Data fusion is defined as "the process of fusing multiple records representing the same real-world object into a single, consistent, and clean representation" [BN09]. In the context of the Semantic Web, previous works on data fusion can be categorized into two classes – *heuristic-based* and *probability-based*.

Heuristic-based Methods. Schultz et al. introduced '*LDIF*', that uses user provided heuristics to find duplicate real-world entities [SMI+10]. Mendes et al. proposed '*Sieve*', which resolves conflicts in Linked Data from different sources by selecting one value for each property based on quality measures such as recency and frequency [MMB12, BB14]. '*ODCleanStore*' provides heuristic-based mechanisms such as max frequency for resolving conflicts during the fusion of Linked Data [KMD+12, MK12]. One of the limitations of such heuristics-based approaches is that they rely on the observation of a specific dataset, which is often not generalizable for other datasets. Furthermore, the used heuristics usually focus on a single aspect of the quality, e.g. recency or frequency, while the quality of a resource is typically influenced by multiple factors to varying degrees.

Probability-based Methods. Zhao et al. [ZRGH12] proposed an unsupervised probabilistic graphical model to infer true records and source quality based on the false-positives and false-negatives of the data source. Dong et al. [DGH+14b] introduced data fusion techniques which identify true subject-predicate-object triples, that are extracted by multiple extractors and originate from multiple sources [DGH+14a]. Pochampally et al. proposed to use joint precision and joint recall to indicate the correlation between sources in order to penalize the copying between sources [PDSD+14]. In later work, the authors proposed a probabilistic model to compute the *Knowledge-Based Trust (KBT)*, i.e., a score for measuring the trustworthiness of the resources [DGM+15]. KBT focuses on the general quality of a resource, and is computed based on the overlap of the extracted data and the knowledge base.

Whereas previous works focus only on the correctness of the source and assign equal weights to all the facts from the same source, in contrast, we not only consider the source quality but also features of the predicates, facts and entities. Hence, distinct facts from the same source are classified differently, depending on multiple feature dimensions. Thus, through a more fine-grained classification, our data fusion approach is able to improve both precision and recall. Another difference is that our query-centric data fusion approach does not require the fusion of the entire dataset after partial changes to the corpus, but can be applied iteratively over specific subsets.

Our recently published work presents an entity summarization approach that retrieves entities from WDC and selects distinct facts to build entity descriptions based on clustering [YGZ+16]. Additional recent work [YGFD17] proposes a data fusion approach focused on ensuring the correctness of facts obtained from noisy entity descriptions in Web markup. While the focus of the former work was on deduplication, the main focus of the latter was correctness. In contrast, this work proposes a com-

plete two-step pipeline aiming at obtaining correct and non-redundant facts which augment existing KBs.

3.2 Motivation & Approach

3.2.1 Motivation

Previous works [SGY+16, DTY+17, YFGD16] have investigated the nature of several type-specific subsets of Web markup, namely bibliographic data, metadata about learning content, books and movies. These works assess markup data on several dimensions, such as data quality, the source distribution and the schema usage. Results show the complementary nature of markup data when compared to traditional knowledge bases, where the extent of additional information varies strongly between types.

For a preliminary analysis of *DBpedia*, *Freebase* and *Wikidata*, we randomly select 30 Wikipedia entities of type *Movie* and *Book* and retrieve the corresponding entity descriptions from all three KBs. We select the 15 most frequently populated properties for each type and provide equivalence mappings across all KB schemas as well as the *schema.org* vocabulary manually[3]. Since all vocabulary terms and types in the following refer to *schema.org*, prefixes are omitted. Figure 3.1 shows the proportion of instances for which the respective properties are populated. We observe a large amount of empty slots across all KBs for most of the properties, with an average proportion of missing statements for books (movies) of 49.8% (37.1%) for DBpedia, 63.8% (23.3%) for Freebase and 60.9 % (40%) for Wikidata.

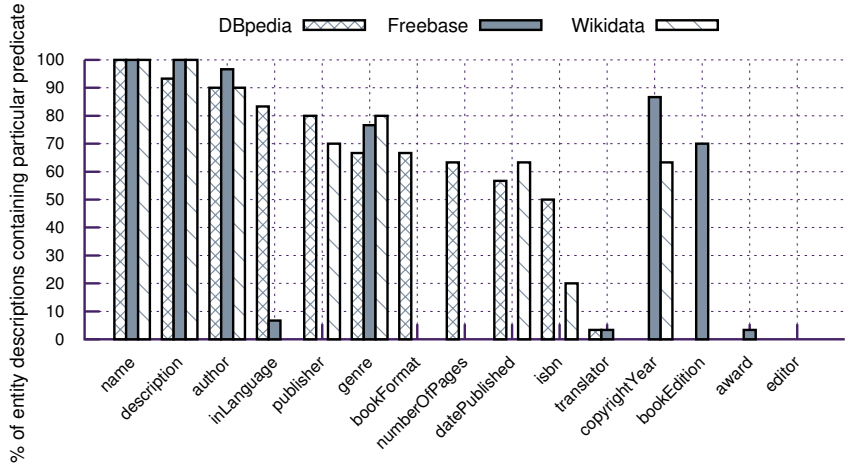
In addition, coverage varies heavily across different properties, with properties such as *editor* or *translator* being hardly present in any of the KBs.

Tail entities/types as well as time-dependent properties which require frequent updates, such as the *award* of a book, are prevalent in markup data [MRP16], yet tend to be underrepresented in structured KBs. Hence, markup data lends itself as a data source for the KBA task. However, given the specific characteristics of markup data [YFGD16], namely the large number of coreferences and near-duplicates, the lack of links and the variety of errors, data fusion techniques are required which are tailored to the specific task of KBA from Web markup.

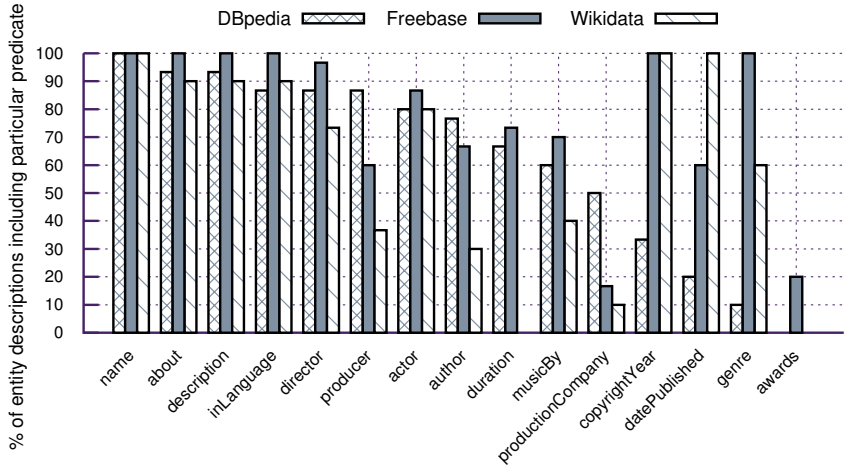
3.2.2 Problem Definition

For the purpose of this work, an *entity description* is considered a semi-structured representation of an actual *entity*, where the latter is either a physical (e.g. a person) or an abstract notion (e.g. a category or theory). Entity descriptions which represent

[3]The mappings are online at: http://l3s.de/~yu/knowmore/



(a) `Book`



(b) `Movie`

Figure 3.1 Proportion of book and movie instances per KB that include selected popular predicates.

the same entity are considered to be *coreferences*.

In particular, our work is concerned with entity descriptions extracted from structured Web markup of Web documents. We refer to such a dataset as M, where the WDC dataset is an example. Data in M consists of entity descriptions e_i, each consisting of a set of RDF quads, i.e. a set of $\langle s, p, o, u \rangle$ quadruples which are referring to entities. The elements $\langle s, p, o, u \rangle$ of the quadruple represent subject, predicate, object and the URL of the document from which the triple $\langle s, p, o \rangle$ has been extracted, respectively. An example of an entity description consisting of 3 quadruples is shown in Table 3.1. Here, for instance, *_:node1* corresponds to subject s, $\langle$*http://schema.org/author*$\rangle$ corresponds to predicate p, *"Evelyn Waugh"* corresponds

to object o and $\langle http://example.url \rangle$[4] corresponds to u, i.e. the document where the observed triples were extracted from. In later sections, for the sake of clarity and readability, we simplify the representation of quads by omitting the namespace and certain formattings such as double quotes.

Table 3.1 Example of an entity description of entity "*Brideshead Revisited*" (of type *Book*) extracted from Web markup.

Quadruples
_:node1 ⟨http://schema.org/author⟩ "Evelyn Waugh" ⟨http://example.url⟩
_:node1 ⟨http://schema.org/bookFormat⟩ "Paperback" ⟨http://example.url⟩
_:node1 ⟨http://schema.org/publisher⟩ "Back Bay Books" ⟨http://example.url⟩

There exist $n \geq 0$ subjects $\{s_1, s_2, ..., s_n\}$, and consequently, n entity descriptions $e_i = \langle s_i, p_i, o_i \rangle \in E$ which represent a particular query entity q in M. Here, E is the set of all entity descriptions which (co)refer to entity q. We define a property-value pair $\langle p, o \rangle$ describing the entity q as a fact of q. Note that we explicitly consider multi-valued properties, i.e. a particular predicate p might be involved in more than one fact for a particular entity q.

We define the task of augmenting a particular entity description e_q, representing a query entity q within a particular KB from data in a markup corpus M as follows:

Definition 1 *KBA task:*

For a query entity q that is represented through an entity description e_q in a KB, we aim at selecting a subset F_{nov} from M, where each fact $f_i \in F_{nov}$ represents a valid fact which augments the entity description e_q for q.

Note that F_{nov} represents the final output of the KnowMore pipeline. We consider a fact valid for augmentation, if it meets the following criteria:

- A fact is *correct* with respect to query entity q, i.e. consistent with the real world regarding query entity q according to some ground truth (Section 3.5).
- A fact represents *novel*, i.e. not duplicate or near-duplicate, information with regard to the entity description e_q of q in a given KB.
- The predicate p_i of fact $\langle p_i, o_i \rangle$ should already be reflected in a KBs given schema.

As an illustrative example, let q be the book *Brideshead Revisited*. In a given KB, such as DBpedia, there is an entity description[5] e_q which represents q. From the Web

[4] For simplicity, we use http://example.url to represent the original URL: http://www.abebooks.com/products/isbn/9780316926348/9697700088.

[5] http://dbpedia.org/resource/Brideshead_Revisited

markup corpus (M), we can extract a set of coreferring entity descriptions E representing the query entity q, that is, the book *Brideshead Revisited*. An example entity description $e_i \in E$ consists of 3 triples $\{\langle$*_:node1, author, Evelyn Waugh*$\rangle$, $\langle$*_:node1, datePublished, 1940*$\rangle$, $\langle$*_:node1, isbn, 9781904605577*$\rangle\}$. From all the facts (e.g. $\langle$*isbn, 9781904605577*$\rangle$) in E, we aim at selecting the ones that are valid, according to the previous definition, for augmenting the KB at hand.

3.2.3 Approach Overview

Our approach (*KnowMore*) for addressing the KBA problem defined above consists of two steps, namely (i) entity matching, and (ii) data fusion. We introduce the intuition behind each step below and describe the actual method in the following sections.

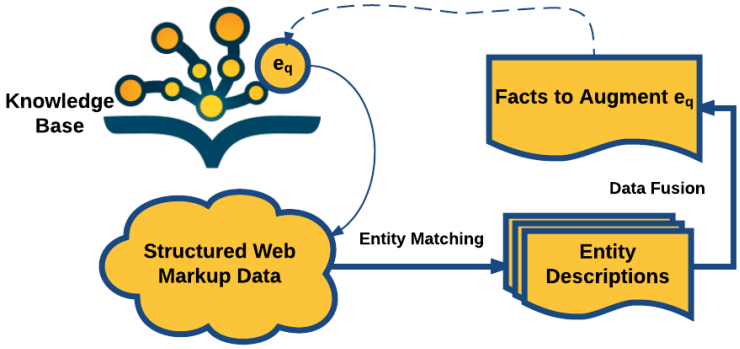


Figure 3.2 Overview of pipeline.

Entity matching. The first step, $KnowMore_{match}$, aims at obtaining candidate facts by collecting the set E of coreferring entity descriptions $e_i \in E$ from M which describe q and corefer to the entity description e_q in a given KB. We use a three step approach in order to efficiently achieve high accuracy results.

- Data cleansing to improve general data quality.
- Blocking with standard BM25 entity retrieval on the value of property *name* of all indexed entity descriptions to reduce the search space. This step results in a set of candidate entity descriptions E_0 that potentially describe the same entity as e_q.
- Validation of each entity description $e_i^0 \in E_0$ in the result of the blocking step using supervised classification on the similarity vector between e_i^0 and e_q.

Hence, we retrieve the set E containing candidate entity descriptions represented through facts $f \in F$ that potentially describe q.

Data fusion. During the data fusion step, $KnowMore_{class}$, we aim at selecting a subset $F_{nov} \subset F$ that fulfills the criteria as listed in Section 3.2.2. More specifically, we introduce *data fusion* techniques based on supervised classification to ensure the correctness (Section 3.4.1) and two deduplication steps to ensure novelty (Section 3.4.2), namely deduplication with respect to M ($KnowMore_{ded}$) and deduplication with respect to the KB ($KnowMore_{nov}$). The latter is a prerequisite for the KBA task.

For clarity, we summarize the notations used for identifying each step of the $KnowMore$ pipeline and the corresponding in- and outputs in Table 3.2.

Table 3.2 Summary of involved steps.

Step	Notation	Input	Output
Entity matching	$KnowMore_{match}$	q, M	F
Data fusion - correctness	$KnowMore_{class}$	F	F_{class}
Deduplication with respect to M	$KnowMore_{ded}$	F_{class}	F_{ded}
Deduplication with respect to KB	$KnowMore_{nov}$	F_{ded}	F_{nov}

We describe each step of the approach in the following two sections in detail.

3.3 Entity Matching

The entity matching step ($KnowMore_{match}$) aims at detecting a set of candidate entity descriptions $e_i \in E$ with $E \subset M$ which are likely to be coreferences of a given KB entity description e_q. We apply three steps in $KnowMore_{match}$, namely cleansing, blocking and matching. These steps are applied over all entities which are to be augmented.

3.3.1 Data Cleansing

This initial data cleansing step aims at (i) resolving object references and (b) fixing common errors [MP15] to improve overall usability of the data.

While *schema.org* property range definitions are not bound in a strict way but constitute mere recommendations, previous studies [DTY+17] observe a strong tendency towards statements which refer to literals rather than objects, i.e. URIs. For instance, within a markup corpus of 44 million quads, 97% of transversal properties referred to literals rather than URIs/objects, despite the fact that only 64% of quads involved properties where *schema.org* recommends literals as property range [DTY+17]. Given this prevalence of literals in Web markup and the need to homogenise entity descriptions for further processing, we resolve object references into literals by replacing object URIs with the labels of the corresponding entity.

In addition, based on earlier work [MP15] which studied common errors in Web markup, we implement heuristics and apply these to E as a cleansing step thereby improving the quality of the data. In particular, we implemented the heuristics as proposed in [MP15] to:

- **Fix wrong namespaces.** Most namespace issues seem due to typing errors, e.g. lacking a slash or using *https://* instead of *http://*. Another reason is the misuse of the upper/lower-cases in a case-sensitive context, e.g. use of *Schema.org* instead of the valid term *schema.org*.
- **Handle undefined types and properties.** The use of undefined types and properties is frequent in Web markup data. Some of the undefined types exist due to typos and the misuse of the upper/lower-cases, e.g. the use of *creativework* for the intended type *CreativeWork*, where simple heuristics can be applied to resolve these issues.

Applying these heuristics improves the performance of the subsequent step by providing a wider and higher quality pool of candidates.

3.3.2 Blocking

Blocking is typically used as a pre-processing step for entity resolution to reduce the number of required comparisons by placing potentially relevant entity descriptions, i.e. potential coreferences, into the same block so that the entity resolution algorithm is applied to entity descriptions within the same block only [CES15].

Related work [TFDCM16] shows that string comparison between labels of markup entities is an efficient way for obtaining potential coreferences, whereas the Lucene BM25 retrieval approach has been used successfully by previous works on entity resolution as summarized by [CES15].

Therefore, we implement the blocking step through entity retrieval using the BM25 retrieval model, i.e. a probabilistic ranking function used to rank matching documents according to their relevance to a given search query, to reduce the search space. We created an index for each type-specific subset using Lucene, and then use the *label* of e_q to query the field *name* within a type-specific index. Hence, queries for a specific type/label-combination which represents q result in a set of candidate entity descriptions $e_i^0 \in E_0$ that potentially describe the same entity as e_q.

Given that the *name*[6] property is one of the most frequently populated properties for the considered entity types, i.e. 90.2% of entity descriptions of type *Book* and 86.8% of entity descriptions of *Movie* are annotated with a name, entity retrieval on the name/label ensures comparably high recall during the blocking step.

[6] http://schema.org/name

For instance, considering the query "*Brideshead Revisited*" (of type *Book*), as part of the blocking step we query the Lucene index and obtain 1,657 entity descriptions consisting of 15,940 quads. An excerpt of the result set is shown in Table 3.3.

Table 3.3 Excerpt from the result set (1,657 entity descriptions in total) for the query "*Brideshead Revisited*" (of type *Book*) after blocking.

Subject	Predicate	Object
_:node1	author	Evelyn Waugh
_:node1	datePublished	1940
_:node1	isbn	9781904605577
_:node2	author	Waugh, Evelyn
_:node2	publisher	Back Bay Books.
_:node3	author	Roger Parsley
_:node3	publisher	Samuel French Ltd

3.3.3 Entity Matching

When attempting to match entities, one can build on the observation that particular property-value pairs can be considered near-unique identifiers for a specific entity, so-called pseudo-key properties. For instance, *taxID* can be considered one of the pseudo-key property for instances of type *Person*, and *isbn* for instances of type *Book*. However, as studied by Meusel et al. [MRP16], resolving coreferences simply through pseudo-key properties does not produce sufficient results when applied on sparsely described and heterogeneous entity descriptions obtained from Web markup.

Thus, we adapt the entity matching approach described in [RM16] to filter out noise in E_0, for instance, entity descriptions which are not relevant to q but fetched through the initial blocking step due to ambiguous labels. Our matching approach builds on the assumption that the importance of distinct properties differs when computing entity similarity, with pseudo-key properties being the most decisive ones. For example, instances of type *Book* which share the same *author* have a higher probability to be equivalent than books which share the same *bookFormat*.

In order to compute the similarity for each property, we consider all properties as attributes of the feature space $\overrightarrow{A} = \{a_1, a_2, ..., a_n\}$, so that each entity description e can be represented as a vector of values $\overrightarrow{v} = \{o_{a1}, o_{a_2}, ..., o_{a_n}\}$ which represent the objects of the considered $\langle p, o \rangle$ tuples. We construct a similarity vector $\overrightarrow{sim}(\overrightarrow{v^{KB}}, \overrightarrow{v})$ between e_q and each entity description $e_i^0 \in E_0$ as in Equation 3.1.

$$\overrightarrow{sim}(\overrightarrow{v^{KB}}, \overrightarrow{v}) = \{\lambda_{a_1}, \lambda_{a_2}, ..., \lambda_{a_n}\} \tag{3.1}$$

$$\lambda_{a_i} = sim(o_{a_i}^{KB}, o_{a_i}) \tag{3.2}$$

In order to compute $sim(o_{a_i}^{KB}, o_{a_i})$, we employ datatype-specific similarity metrics, i.e., we implemented one similarity measure for each *schema.org* datatype[7], and automatically select the appropriate metric. Specifically, for 1) *Text*, we employ cosine similarity, for 2) *Number* or *Boolean* attributes, $sim(o_{a_i}^{KB}, o_{a_i})$ equals to 1 if $o_{a_i}^{KB}$ is the same as o_{a_i}, and 0 otherwise, for 3) *Time* or *DateTime*, we first unify different formatting styles with the *java DateTimeFormatter*[8] class, and then split values into separate date parts (i.e. year, month, date), where $sim(o_{a_i}^{KB}, o_{a_i})$ is 0 if there is a conflict any of the observed parts. For example, *1990 April* and *May 10th* would constitute a conflict as the month unit is present in both strings, but the values are different. This indicates that these two strings cannot possibly represent the same date. Otherwise, i.e., in cases where no direct conflict is observed, we compute the Jaccard similarity between both triples consisting of the value of $\langle year, month, date \rangle$ as a metric of the overlapping semantics. For instance, the dates "2017 Oct 1st" and "2017 Oct" have a similarity of 0.667 as they have 2 common units out of 3. The java implementation of this step can be found online[9].

We then train a supervised classification model, to make the decision whether or not e_i^0 is a match for e_q. We experimented with several state-of-the-art classifiers (SVM, Logistic Regression and Naive Bayes). Since Naive Bayes achieves a $F1$ score that is 0.08 higher than the best SVM (linear kernel), and 0.123 higher than the Logistic Regression (LR), throughout the remaining of this work we rely on a trained Naive Bayes classifier unless otherwise stated. More details about classifier performance are provided in Section 3.6.1. The classification and clustering implementation in our approach is built on top of the Java-ML toolkit[10]. The training data is described in Section 3.5.2.

The final result of the entity matching step is the set of coreferring entity descriptions $e_i \in E$ which constitute candidate facts $f_i \in F$ for the following steps.

Returning to our running example "*Brideshead Revisited*", after removing the unmatched entity descriptions from the blocking result through the entity matching step, there are 44 matched entity descriptions remaining in the result set. Some examples are shown in Table 3.4, where, for instance, *_:node3* had been removed since it refers to the stage play rather than the book and does not match entity q.

3.4 Data Fusion

This step aims at fusing candidate entity descriptions in E by detecting the correct and novel facts $f_{nov} \in F_{nov}$ with $F_{nov} \subset F$ to augment e_q.

[7] http://schema.org/DataType
[8] java.time.format.DateTimeFormatter
[9] http://l3s.de/~yu/knowmore/
[10] http://java-ml.sourceforge.net/

Table 3.4 Excerpt from result set (44 entity descriptions in total) for the query, "*Brideshead Revisited*" (of type *Book*) after entity matching.

Subject	Predicate	Object
_:node1	author	Evelyn Waugh
_:node1	datePublished	1940
_:node1	isbn	9781904605577
_:node2	author	Waugh, Evelyn
_:node2	publisher	Back Bay Books.

3.4.1 Correctness - Supervised Classification

Table 3.5 Features for supervised data fusion from markup data.

Category	Notation	Feature description
Source level	t_1^r, t_2^r, t_3^r	Maximum, minimum, average PageRank score of the PLDs containing fact f
	t_4^r, t_5^r, t_6^r	Maximum, minimum, average percentage of common errors [MP15] of the PLDs containing fact f
	t_7^r, t_8^r, t_9^r	Maximum, minimum, average precision (based on training data) of the PLDs containing fact f
Entity level	t_1^e, t_2^e, t_3^e	Maximum, minimum, average size (number of facts) of e_i containing f
Property level	t_1^p	Predicate term
	t_2^p	Predicate frequency in F
	$\dagger t_3^p$	Amount of clusters of predicate p
	$\dagger t_4^p$	Average cluster size of predicate p
	$\dagger t_5^p$	Variance of the cluster sizes of predicate p
Fact level	t_1^f	Fact frequency in F
	$\dagger t_2^f$	Normalized cluster size that f belongs to

†-features extracted based on clustering result

The first step ($KnowMore_{class}$) aims at detecting correct facts by learning a supervised model that produces a binary classification for a given fact $f \in F$ into one of the labels {*'correct'*, *'incorrect'*}. For the classification model, we have experimented with several different approaches, namely Naive Bayes classification, SVM with different kernel functions, Logistic regression, kNN with varying k's and Random Forest. We rely on a Naive Bayes classification since our experiments have shown superior performance over the second best performing approach Logistic Regression with an increase in $F1$ score of 0.016, and over SVM (linear kernel) with an increase in $F1$ score of 0.044. Details about the performance of the three best performing approaches

(Naive Bayes, SVM and Logistic regression) are provided in Section 3.6.2. We introduce the features used for our supervised learning approach in Table 3.5 and describe them in detail below. Through an initial data analysis step, all features have been identified as potential indicators of fact correctness.

While we aim to detect the correctness of a fact, we consider characteristics of the *source*, that is the Pay-Level-Domain (PLD, i.e. the sub-domain of a public top-level domain, which Website providers usually pay for), from which a fact originates, the *entity description*, the *predicate* term as well as the *fact* itself. The four different categories are described below.

Source level. As has been widely studied in previous works, source quality is an important indicator for data fusion [ZRGH12, PDSD+14, DGM+15]. Features $t^r_1, ..., t^r_3$ are generated from the PageRank score as an authority indicator of the PLD from which a fact is extracted, assuming a higher PageRank indicates higher authority and hence quality. Based on the intuition that more errors across the markup from of respective PLD indicate a higher potential of this PLD to provide incorrect facts, we consider the rate of common errors detected based on previously identified heuristics [MP15] to compute features $t^r_4, ..., t^r_6$. Finally, we use precision of a PLD computed based on our ground truth (Section 3.5.2) as quality indicators in features $t^r_7, ..., t^r_9$.

Entity level. Based on the data analysis, entity descriptions containing a large number of facts are usually of higher quality. Thus, we use the size of entity descriptions, reflected through features $t^e_1, ..., t^e_3$, as additional indicator of quality.

Property level. The quality of facts strongly varies across predicates, as identified in previous studies [YFGD16, MRP16], with some properties being more likely to be part of a correct fact than others. One example of a predicate often included in incorrect statements is *datePublished* of a movie, that is often mistakenly used to describe the publishing time of the Web document. Following this observation, we extract features $t^p_1, ..., t^p_5$ to consider characteristics of the involved predicate terms, such as their frequency.

Given that our candidate set contains vast amounts of near-duplicate facts, we approach the problem of identifying semantically equivalent statements through clustering of facts which use varied surface terms for the same or overlapping meanings. We employ the X-Means algorithm [PM+00], as it is able to automatically determine the number of clusters. This clustering step aims at grouping or canonicalizing different literals or surface forms for specific object values. For instance, *Tom Hanks* and *T. Hanks* are equivalent surface forms representing the same entity. To detect duplicates and near-duplicates, we first cluster facts that have the same predicate p into n clusters $(c_1, c_2, \cdots, c_n) \in C$. In this way, considering string similarity, we can canonicalize equivalent surface forms. The performance of the clustering on removing near duplicates is discussed in Section 3.6.3. Another challenge considered here is the cardinality of predicates. Depending on the predicate, the number of potentially correct statements varies. For example, *actor* is associated with multiple

values, whereas *duration* normally has only one valid statement. This is reflected in the cluster amount n for a given predicate (t_3^p). The intuition behind feature t_4^p is that the average size of clusters is an indicator of the frequency of facts in p which usually correlates with the quality. Feature t_5^p is extracted based on the observation that in most cases, wrong facts have lower frequency than average, thus the variance is larger if there are wrong facts among the facts of p.

Fact level. Fact frequency [MMB12] has been used in previous data fusion works and is shown to provide efficient features for determining the correctness of facts. Based on these insights, we extract features t_1^f and t_2^f. We consider the size of a cluster as feature t_2^f indicating the frequency of a fact, where the normalized size of cluster c_i is $|c_i|/\sum_{j=1}^{n}|c_j|$.

From the computed features we train the classifier for classifying the facts from F into the binary labels {*'correct'*, *'incorrect'*}. More details about the training and evaluation through 10-fold cross-validation are presented in Section 3.6.2. The *'correct'* facts form a set F_{class} that is the input for the next steps.

Again returning to our running example, after removing wrong facts from the candidate facts, such as *datePublished: 1940* through the classification step, we obtain 37 correct facts in the result set for the query *Brideshead Revisited, type:(Book)*. An excerpt of the resulting facts are shown in Table 3.6.

Table 3.6 Excerpt from result set (37 distinct correct facts) for query "*Brideshead Revisited*" (of type *Book*) for $KnowMore_{class}$.

Class	Predicate	Object
correct	s:author	Evelyn Waugh
correct	s:isbn	9781904605577
correct	s:author	Waugh, Evelyn
incorrect	datePublished	1940
correct	s:publisher	Back Bay Books.

3.4.2 Novelty

A fact f is considered to be *novel* with respect to the KBA task, if it fulfills the conditions: i) is not duplicate with other facts selected from our source markup corpus M, ii) is not duplicate with any facts existing in the KB. Each of these two conditions corresponds to a deduplication step.

Deduplication with respect to M ($\mathbf{KnowMore_{ded}}$). As introduced in Section 3.4.1, we detect near-duplicates via clustering. For each predicate p, all the facts $f = \langle p, o_i \rangle$ corresponding to p are clustered into n clusters $\{c_1, c_2, \cdots, c_n\}$. Each cluster $c_i, i = 1, ..., n$ contains a set of near-duplicates. To fulfill i), we select only

one fact from each cluster by choosing the fact that is closer to the cluster's centroid. This results in the fact set F_{ded} that is the input for next deduplication step.

Deduplication with respect to KB (KnowMore$_{\text{nov}}$). We compute the similarity $sim(f_i, f_{KB})$ between a fact f_{KB} in a respective KB for a particular predicate p and a fact f_i for the same (mapped) predicate p in F_{ded} with the datatype-specific similarity metrics as introduced in Section 3.3. If $sim(f_i, f_{KB})$ is higher than a threshold τ, we remove the fact along with its near-duplicates, i.e. the facts in the same cluster from the candidate set F_{nov}. We explain τ and its configuration during the experimental Section 3.5.3. The facts selected from F_{nov} in this step are the final result for augmenting the KB.

In our running example, the fact *author: Waugh, Evelyn* is removed during the deduplication with regard to M as it is a duplicate of fact *author: Evelyn Waugh*, which has been selected as more representative.

With respect to the KBA task, consider the augmentation of the example entity "*Brideshead Revisited*" (of type *Book*) as illustrated in Table 3.7. The example facts #2 and #3 would be valid results of the KBA task for DBpedia since they are *novel*, while only fact #2 is a valid augmentation for Freebase and Wikidata as it is the only fact that is novel.

Table 3.7 Novelty of correct, distinct facts with regard to KBs for the query "*Brideshead Revisited*" (of type *Book*).

ID	Fact	DBpedia	Wikidata	Freebase
1	author, Evelyn Waugh	✗	✗	✗
2	isbn, 9781904605577	✔	✔	✔
3	publ., Back Bay Books	✔	✗	✗

Note that our deduplication step considers and supports multi-valued properties. By relying on the clustering features, computed during the fusion step, we select facts from multiple clusters (corresponding to multiple predicates) as long as they are classified as correct. As documented by the evaluation results (Section 6.4), this does not negatively affect precision while improving recall for multi-valued properties.

3.5 Experimental Setup

3.5.1 Data

Dataset. We use the WDC2015 dataset[11], where we extracted 2 type-specific subsets consisting of entity descriptions of the *schema.org* types *Movie* and *Book*. Initial experiments indicated that these types are well reflected in the WDC2015 datasets,

[11] http://webdatacommons.org/structureddata/index.html#toc3

and at the same time, their facts are comparably easy to validate manually when attempting to label a ground truth. The *Movie* subset consists of 116,587,788 quads that correspond to 23,334,680 subjects/nodes, and the *Book* subset consist of 174,459,305 quads and 34,655,078 subjects.

Entities & KBs to Augment. As input for the KBA task, we randomly select 30 entities from Wikipedia for each type *Book* and *Movie*. We evaluate the performance of our approach for augmenting entity descriptions of these 60 entities obtained from three different KBs: DBpedia (*DB*), Freebase (*FB*) and Wikidata (*WD*). For DBpedia, we retrieve entity descriptions through the SPARQL endpoint[12] where resource URIs were obtained by replacing the Wikipedia namespace of our selected entities with the DBpedia resource path. URIs of corresponding Freebase and Wikidata entity descriptions are obtained through the *owl:sameAs* links present in DBpedia. Using these URIs, the respective entity descriptions are obtained through the latest available version of Freebase[13] (accessed Sep 30, 2016) and the Wikidata SPARQL endpoint[14].

The full list of entities can be found online[15]. An analysis of the completeness of these obtained entity descriptions is shown in Figure 3.1 in Section 3.2.1.

Properties to Augment. To simplify the schema mapping problem between WDC data and the respective KBs while at the same time taking advantage of the large-scale data available in our corpus, we limit the task to entities annotated with the *http://schema.org* ontology for this experiment. Previous works have shown that *schema.org* is the only vocabulary which is consistently used at scale [MRP16]. We manually create a set of schema mappings that maps the *schema.org* vocabularies to the *DB, FB, WD* vocabularies. For this, we first select all the *schema.org* predicates appearing in F. We identify the ones that have equivalent properties within all involved vocabularies and create equivalence mappings (*owl:equivalentProperty*). The list of predicates and the mapping statements can be found online[16].

3.5.2 Ground Truth & Metrics

This section describes the ground truth used for training and testing together with the evaluation metrics used for assessing performance in different tasks.

Ground Truth via Crowdsourcing

Entity Matching. We used crowdsourcing to build a ground truth by acquiring labels for each $e_i \in E$. In each case, crowd workers were presented with the entity

[12] http://dbpedia.org/sparql
[13] http://commondatastorage.googleapis.com/freebase-public/rdf/freebase-rdf-latest.gz
[14] https://query.wikidata.org/sparql/
[15] http://l3s.de/~yu/knowmore/
[16] http://l3s.de/~yu/knowmore/

description e_q, i.e. the Wikipedia page, and entity description $e_i \in E$, and were asked to validate e_i as either *valid*, *invalid* or *insufficient information to judge* with respect to e_q. We deployed the task on CrowdFlower[17], and gathered 5 judgments from distinct workers on each (q, $e_i \in E$) pair. To ensure high quality, we restricted the participation to Level 3 workers alone[18]. In addition to this, we used test questions to flag and reject untrustworthy workers. Workers were compensated at the rate of 6 USD cents per judgment. On average, workers performed with an accuracy of 92% on the test questions, indicating high reliability. The inter-rater agreement between workers was 75% using Krippendorff's Alpha [Kri07], and 89% using pairwise percent agreement (PPA). By applying this process on E_0, we obtain 89 (180) *valid* and 128 (118) *invalid* entity descriptions for *Movie* (*Book*) entities respectively.

Data Fusion - Correctness. Similarly, we used crowdsourcing to build a ground truth for the correctness of facts $f_i \in F$. For the valid entity descriptions in E, we acquire labels for all distinct facts, as either *correct* or *incorrect* with respect to q. We acquired 5 judgments from distinct workers for each entity and corresponding facts through Crowdflower. We used similar quality control mechanisms as in the entity matching task. Workers were compensated at the rate of 6 USD cents per judgment. Workers performed with an accuracy of 95% on the test questions. The inter-rater agreement between workers was 71.1% using Krippendorff's Alpha, and 86.9% using pairwise percent agreement (PPA). This indicates a high reliability of the ground truth. This process results in 371 (out of 456) and 298 (out of 341) *correct* facts for *Movie* and *Book* dataset respectively. Distinct facts were obtained by removing duplicate literals, null values, URLs and the unresolved objects (e.g. *node3* that could not be resolved in the dataset). The ground truth is publicly available[19].

Data Fusion - Novelty. We built corresponding ground truths for validating (i) deduplication performance within M, as well as (ii) novelty with respect to the different *KBs*. Three authors of [YGF+19a] acted as experts and designed a coding frame to decide whether or not a fact is novel. After resolving disagreements on the coding frame on a subset of the data, every fact was associated with one expert label through manual deliberation. We followed the guidelines laid out by Strauss [Str87] during the coding process, which provide guidelines for designing reliable coding frames and carrying out manual coding and are frequently used to design coding frames and conduct qualitative analysis. Distinct facts were obtained by removing duplicate literals, null values, URLs and unresolved objects.

Metrics

We consider distinct metrics for evaluating each step of our approach.

[17] Formerly http:www.crowdflower.com, now https://www.figure-eight.com/

[18] Level 3 workers on CrowdFlower have the best reputation and near perfect accuracy in hundreds of previous tasks.

[19] http://l3s.de/~yu/knowmore/

- $KnowMore_{match}$. To evaluate performance of the matching step, we consider precision P - the percentage of entity descriptions $e_i \in E$ that were correctly matched to e_q, recall R - the percentage of $e_i \in E_0$ that were correctly matched to KB, and the $F1$ score.

- $KnowMore_{class}$. We evaluate the performance of the approaches through standard precision P, recall R and $F1$ scores, based on our ground truth.

- $KnowMore_{ded}$. We evaluate the performance of deduplication with respect to M ($KnowMore_{ded}$) using $Dist\%$ - the percentage of distinct facts within the respective result set. We compare between $Dist\%(F_{ded})$ and $Dist\%(F_{class})$, that is, before and after the deduplication within M.

- $KnowMore_{nov}$. For evaluating the performance of deduplication with respect to a given KB ($KnowMore_{nov}$), we measure the novelty as Nov - the percentage of novel facts - and compare between $Nov(F_{ded})$ and $Nov(F_{nov})$, that is, the novelty before and after this step. We also measure the recall R - the percentage of distinct and accurate facts in F_{ded} that have been selected by $KnowMore_{nov}$ into F_{nov}.

Furthermore, we demonstrate the potential of our approach for augmenting a given KB by measuring the *coverage gain*, which we introduce as a means to measure the capacity of our approach to populate gaps in existing KBs (Section 3.2.1). The coverage gain of predicate p is computed as the percentage of entity descriptions having p populated through the $KnowMore$ approach (i.e. after step $KnowMore_{nov}$) with at least one fact $\langle p, o\rangle$, out of the ones that did not have at least one statement involving property p within the respective KB before augmentation. Note that according to this metric a coverage gain of less than 100% might not necessarily indicate non-optimal recall but might be caused by the non-applicability of attributes to a particular entity. For instance, not all entities of type *Movie* have a value for the property *award*. The result is reported in Section 3.6.4.

3.5.3 Configuration & Baselines

Configuration. We deploy our approach as described in Section 3.3.3 and 3.4. For the entity matching step, we use Lucene for indexing and BM25 retrieval with the Lucene default configuration where $k_1 = 1.2$, $b = 0.75$. For the deduplication with respect to KBs, we report the evaluation result of $KnowMore_{nov}$ using different $\tau = \{0.3, 0.5, 0.7\}$ in Section 3.6.3.

Baselines. We compare ($KnowMore_{class}$) with $PrecRecCorr$ that is proposed by Pochampally et al. [PDSD+14] and $CBFS$ [YGZ+16]. To the best of our knowledge, the $CBFS$ approach is the only available method so far directly geared towards the challenges of markup data, while $PrecRecCorr$ represents a recent and highly

related data fusion baseline. We also present the results of $KnowMore_{match}$ and $KnowMore_{class}$ using different classifiers.

- $KnowMore_{class}$: facts selected based on the $KnowMore_{class}$ pipeline from F, using Naive Bayes as classifier for selecting correct facts.
- $KnowMore_{class}$ (SVM): facts selected based on the $KnowMore_{class}$ pipeline from F, using Support Vector Machine (linear kernel) as the classifier for selecting correct facts.
- $KnowMore_{class}$ (LR): facts selected based on the $KnowMore_{class}$ pipeline from F, using Logistic Regression (LR) as classifier for selecting correct facts.
- $PrecRecCorr$: facts selected based on the approach from candidate set F. We consider each PLD as a source and implemented the *exact solution* as described in the paper. We use the threshold as presented in the paper, i.e. 0.5, to classify facts.
- $CBFS$: facts selected based on the $CBFS$ approach from F. The $CBFS$ approach clusters the associated values at the predicate level into n clusters $(c_1, c_2, \cdots, c_n) \in C$. Facts that are closest to the cluster's centroid of each cluster are selected, provided they meet the following criteria:

$$|c_j| > \beta \cdot \max(|c_k|), c_k \in C \tag{3.3}$$

  where $|c_j|$ denotes the size of cluster c_j, and β is a parameter used to adjust the number of facts. In our experiment, β is empirically set to 0.5, which is the one used by the best-performing setup in the original paper.

3.6 Evaluation Results

In this section, we present experimental results obtained through the setup described in the previous sections.

3.6.1 Entity Matching

As the entity matching step ($KnowMore_{match}$) is a precondition for the subsequent fusion step, we provide evaluation results for this step and compare it to entity descriptions obtained through BM25@k as baseline. The BM25 configuration is the same as used in our approach (Section 3.5.3). Since we obtain the corresponding URIs of Freebase and Wikidata entities through the *sameAs* link in DBpedia, here we present only the result of matching entity descriptions to DBpedia. Table 3.8 shows the evaluation results of the standard precision P, recall R and $F1$ scores.

Table 3.8 Performance of $KnowMore_{match}$ and baselines.

	Movie			Book		
Approach	**P**	**R**	**F1**	**P**	**R**	**F1**
KnowMore$_{match}$	**0.943**	0.742	**0.830**	**0.880**	0.894	**0.887**
KnowMore$_{match}$(SVM)	0.583	**0.870**	0.698	0.824	**0.899**	0.860
KnowMore$_{match}$(LR)	0.627	0.645	0.636	0.821	0.851	0.836
BM25@10	0.659	0.652	0.655	0.325	0.533	0.404
BM25@20	0.592	0.831	0.692	0.219	0.722	0.336
BM25@50	0.406	1.000	0.578	0.124	1.000	0.220

As presented in Table 3.8, our supervised matching approach $KnowMore_{match}$ (using NB as classifier) achieves high $F1$ scores of 0.83 and 0.887 respectively, thereby outperforming the BM25@20 and BM25@50 baselines and providing a sound set of candidates for the subsequent step. One reason for the poor precision of the baseline BM25@k is the introduction of false positives by relying on the value of property *name*, which is inherently ambiguous for the matching task. Since $KnowMore_{match}$ uses BM25 retrieval as a blocking step and then applies a classifier that takes all the properties into consideration, precision is significantly improved. Among different configurations of the $KnowMore_{match}$ approach, the Naive Bayes classifier achieves the highest precision and $F1$ score compared to the Logistic Regression (LR) classifier $KnowMore_{match}(LR)$ and SVM $KnowMore_{match}$(SVM) on both types.

3.6.2 Correctness - Data Fusion

Table 3.9 Performance of $KnowMore_{class}$ and baselines.

	Movie			Book		
Approach	**P**	**R**	**F1**	**P**	**R**	**F1**
KnowMore$_{class}$	**0.954**	0.896	**0.924**	0.880	**0.868**	**0.874**
KnowMore$_{class}$(SVM)	0.845	**0.976**	0.906	0.810	0.799	0.804
KnowMore$_{class}$(LR)	0.894	0.946	0.919	0.889	0.809	0.847
PrecRecCorr	0.924	0.861	0.891	**0.893**	0.48	0.624
CBFS	0.802	0.752	0.776	0.733	0.842	0.784

The results for $KnowMore_{class}$ as well as the baselines are shown in Table 3.9. As shown in the table, our chosen configuration, i.e. using a Naive Bayes classifier ($KnowMore_{class}$) achieves highest $F1$ scores among all the different configurations. Compared to the different configurations of our approach deploying different classification models, the precision is 0.025 (0.089) higher, and the $F1$ score is 0.016 (0.044) higher than using LR (SVM). Potential reasons for the strong performance of the NB classifier could be 1) the assumed independence of features, which is one of the

expectations of NB classifiers and 2) that, compared to other models, NB tends less to overfitting when the amount of training data is limited.

The presented F1 score of the *PrecRecCorr* baseline is the best possible configuration for our given task, where we experimented with different thresholds ([0,1], gap 0.1) as discussed in [PDSD+14] and identified 0.5 experimentally as the best possible configuration. We observe that the $F1$ score of our approach is 0.141 higher than *PrecRecCorr* and 0.119 higher than $CBFS$ on average across datasets. This indicates that our approach provides the most efficient balance between precision and recall across the investigated datasets, when applied to the novel task of data fusion from Web markup. Although, the precision of the baseline approach *PrecRecCorr* is 0.013 higher than the one from $KnowMore_{class}$ on the *Book* dataset, the baseline fails to recall a large amount of correct facts, where the recall of $KnowMore_{class}$ is approximately 0.388 higher. This also is reflected in the average size of entity descriptions obtained through both approaches, where the entity descriptions from *PrecRecCorr* consist of 4.88 statements on average, and the ones from $KnowMore_{class}$ are 8.83, indicating a larger potential for the KBA task.

The reason that *PrecRecCorr* has lower recall compared to $KnowMore_{class}$ is that the *PrecRecCorr* approach relies on the assumption that the facts extracted from high quality sources (i.e. PLDs that provide larger percent of correct facts) are more likely to be true. Although this is a reasonable assumption that results in a high precision result, it penalises facts that originate from sources (PLDs) which contributed wrong facts as part of the training set. This leads to a large number of false negatives. $KnowMore_{class}$, however, considers the source quality (source level features), yet uses several other features in the decision making process and is more robust to the source quality and thus results in higher recall.

The $CBFS$ approach is built based on the intuition that the more frequent a fact is, the higher the chance that it is true. Based on the results, this assumption does not necessarily lead to higher precision compared to the assumption used by the *PrecRecCorr* approach. Since $KnowMore_{class}$ utilizes multiple features extracted from several different dimensions, including fact frequency, the classifier trained on these multi-dimensional features results in better performance.

A more detailed discussion of the potential impact on the KBA task is provided in Section 3.8, investigating the KBA potential beyond the narrow definition of the investigated task of this setup, e.g. by augmenting additional predicates not already foreseen in a given KB schema or to populate KBs with additional entities.

We ran 20 iterations of 10-fold cross validation for different baselines and our approach in order to test the statistical significance of our results, as suggested in [Die98]. We conducted paired T-tests and employed Bonferroni's correction for Type-I error inflation. We found that all comparisons were statistically significant at the 95% confidence interval ($p<0.05$), with some comparisons significant at the 99% confidence interval ($p<0.01$). Tables 3.8 and 3.9 reflect the average values over 20 runs of the corresponding algorithms.

3.6.3 Novelty

This section presents the evaluation results for the deduplication steps introduced in Section 3.4.2.

Diversity. Table 3.10 presents the evaluation result before ($Dist\%$ (F_{class})) and after ($Dist\%$ (F_{ded})) the step $KnowMore_{ded}$.

Table 3.10 Diversity $Dist\%$ before and after deduplication.

Dataset	Dist%($\mathbf{F_{class}}$)	Dist%($\mathbf{F_{ded}}$)
Movie	94.8	**96.1**
Book	82.1	**95.6**

The $Dist\%$ of facts improves by 1.3 percentage points for the *Movie* dataset and by 13.5 percentage points for the *Book* dataset (Table 3.10). The less improvement gain for the *Movie* dataset presumably is due to the nature of the randomly selected *Movie* entities. As these appear to be mostly tail entities, candidate facts in our markup corpus M are fewer and less redundant. Hence, the amount of duplicates and near-duplicates is smaller, reducing the effect of the deduplication step. Deduplication is of particular importance for popular and well-represented entities.

Novelty with respect to KB. The results before (Nov (F_{ded})) and after (Nov (F_{nov})) the deduplication for specific KBs using different similarity thresholds (τ) are presented in Table 3.11.

Table 3.11 Novelty of F_{ded} and F_{nov} with respect to target KBs.

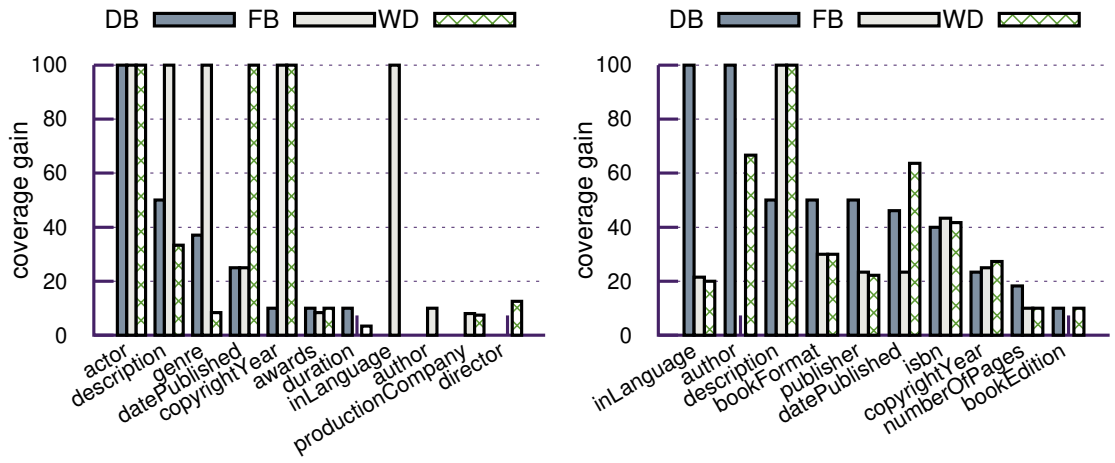
	KB	Nov($\mathbf{F_{ded}}$)	Nov($\mathbf{F_{nov}}, \tau = 0.3$)	Nov($\mathbf{F_{nov}}, \tau = 0.5$)	Nov($\mathbf{F_{nov}}, \tau = 0.7$)
Movie	**DBpedia**	0.631	**0.963**	0.962	0.962
	Freebase	0.527	**0.747**	0.742	0.742
	Wikidata	0.412	**0.929**	0.929	0.897
		R($\mathbf{F_{ded}}$)	**R($\mathbf{F_{nov}}, \tau = 0.3$)**	**R($\mathbf{F_{nov}}, \tau = 0.5$)**	**R($\mathbf{F_{nov}}, \tau = 0.7$)**
	DBpedia	1	0.927	**0.939**	**0.939**
	Freebase	1	0.942	**0.957**	**0.957**
	Wikidata	1	**0.963**	**0.963**	**0.963**
	KB	**Nov($\mathbf{F_{ded}}$)**	**Nov($\mathbf{F_{nov}}, \tau = 0.3$)**	**Nov($\mathbf{F_{nov}}, \tau = 0.5$)**	**Nov($\mathbf{F_{nov}}, \tau = 0.7$)**
Book	**DBpedia**	0.736	**0.962**	0.929	0.92
	Freebase	0.639	**0.915**	0.846	0.825
	Wikidata	0.705	**0.944**	0.933	0.923
		R($\mathbf{F_{ded}}$)	**R($\mathbf{F_{nov}}, \tau = 0.3$)**	**R($\mathbf{F_{nov}}, \tau = 0.5$)**	**R($\mathbf{F_{nov}}, \tau = 0.7$)**
	DBpedia	1	0.826	0.848	**0.870**
	Freebase	1	0.833	**0.846**	**0.846**
	Wikidata	1	0.791	0.814	**0.837**

Since our approach is not aware of the total number of novel facts for a particular

entity description on the Web a priori, in this evaluation, we consider all the novel facts in F_{ded} as the gold standard, and compute the recall of F_{nov} after applying the $KnowMore_{nov}$ accordingly. We evaluate the performance of $KnowMore_{nov}$ using τ in {0.3, 0.5, 0.7} since (1) 0.5 is widely used as threshold for identifying duplicate text using cosine similarity, (2) for computing DateTime similarity between facts, the possible similarities according to our metric are: {0, 0.33, 0.5, 0.67, 1}. These 3 selected thresholds are most influential on the selection of DateTime facts, thus can produce most conclusive results for showing the trade-off between novelty and recall. As shown in Table 3.11, even though there is a trade-off between *novelty* and *recall*, different values of τ do not have a strong influence on the evaluation metrics. One of the reasons is that, a large proportion of facts have non-literal (e.g. numeric) values. While our datatype-specific similarity computes a binary (0 or 1) score in these cases, it is not influenced by the selection of τ.

Consider $\tau = 0.5$ as an example. The $KnowMore_{nov}$ step improves novelty by 0.282 on average across datasets and KBs compared to the result of the $KnowMore_{ded}$ step (F_{ded}), and is able to recall over 90% of the novel facts. Our final result F_{nov} shows a novelty of over 90% on average, what translates to a minor amount of near-duplicates and a sufficient novelty for augmenting the target KBs.

3.6.4 Coverage Gain



(a) Movie **(b)** Book

Figure 3.3 Proportion of augmented entity descriptions with $KnowMore$. Only predicates which were augmented in at least one KB are shown.

This section discusses the coverage gain, as an indicator of the $KnowMore$ performance in the particular KBA task described in Section 3.2.2. Since the addressed task is fairly narrow, i.e. dealing with the augmentation of a selected set of attributes and

entities existing in all three investigated KBs, we discuss the overall KBA potential of Web markup data beyond our ground truth dataset in Section 3.8.1.

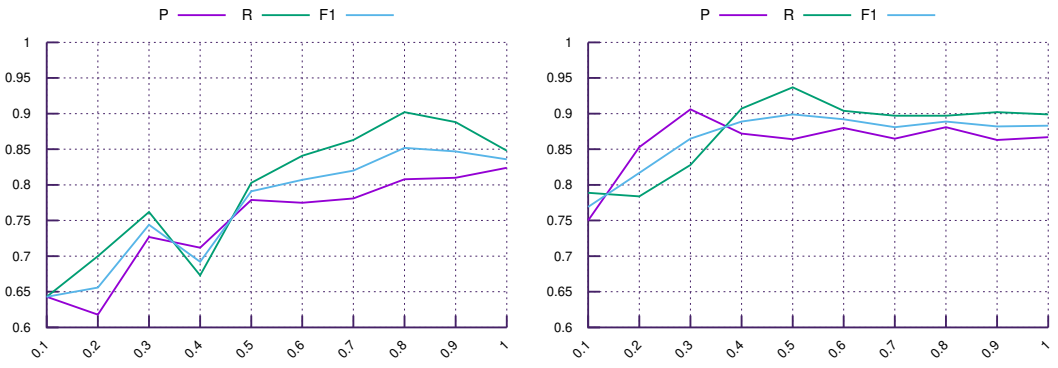
Figure 3.3 shows the coverage gain on the previously empty slots as shown in Figure 3.1 per predicate and KB for our selected 30 entities (per type). Based on the evaluation result, the *KnowMore* pipeline shows a coverage gain of 34.75% on average across different properties for DBpedia, 39.42% for Freebase and 36.49% for Wikidata. We observe that the obtained gain varies strongly between predicates and entity types, with a generally higher gain for book-related facts. For instance, within the *Movie* case, for property *actor* we were able to gain 100% coverage in both DBpedia and Freebase, while the property *award* shows a coverage gain of 10% or less for all three KBs. Reasons behind low coverage gain for a particular property are 2-fold: 1) the lack of data in the Web markup data corpus, and 2) the lack of true facts in the real world for a particular attribute, e.g. only a small proportion of movies have won an award. On average, we obtained 2.8 (6.8) facts for each movie (book) entity in our experimental dataset.

For a more thorough discussion of the KBA potential of the *KnowMore* approach, i.e. on tasks beyond our ground truth dataset, we refer the reader to Section 3.8.1.

3.7 Evaluation of Generalisation Potential

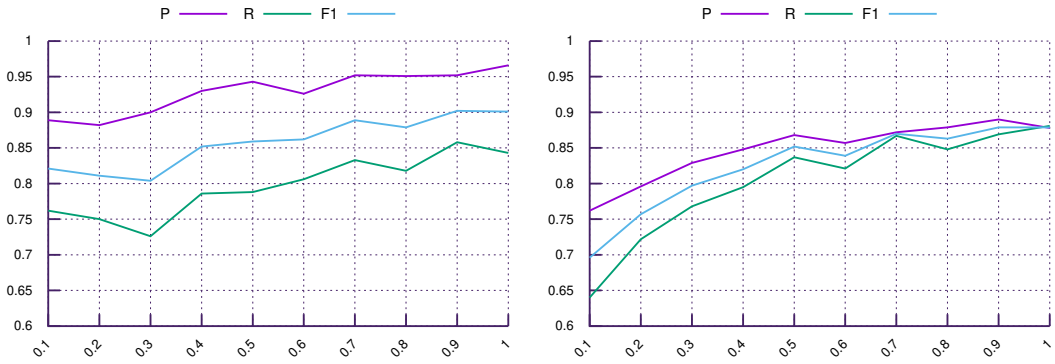
As introduced earlier, our approach has been trained on two specific types (*Book* and *Movie*). The intuition behind this choice is that (i) different properties have varied contribution on different types when computing the similarity between entity descriptions in the entity matching step and (ii) particular features such as predicate term t_1^p increase the feature space when introducing new types and associated properties. To this end, we have restricted our experiments to particular types to reduce the required training data. However, given the type-agnostic nature of most features, it seems reasonable to anticipate comparable performance even when applying our approach across types.

In this section, we evaluate the generalisation of the *KnowMore* approach with respect to two aspects: 1) the scale of training data required for the supervised fusion step, 2) the performance of our approach when trained with cross-type training data, as opposed to type-specific sets.



(a) *Movie* **(b)** *Book*

Figure 3.4 P, R and F1 score using different size of the training data for $KnowMore_{match}$. X-axis shows the percent of training data, Y-axis shows the P/R/F1 value.



(a) *Movie* **(b)** *Book*

Figure 3.5 P, R and F1 score using different size of training data for $KnowMore_{class}$. X-axis shows the percent of training data, Y-axis shows P/R/F1 value.

3.7.1 Scale of required Training Data

As described in Section 3.5.2, our ground truth consists of labels for 217 (298) entity descriptions for *Movie* (*Book*) entities for the entity matching step $KnowMore_{match}$, and 456 (371) facts of *Movie* (*Book*) entities for the data fusion step $KnowMore_{class}$. The experimental evaluation in Section 3.6 is based on the averaged P/R/F1 scores from a 10-fold cross validation (90% training and 10% testing). To evaluate how performance is affected by the scale of the training data, we have conducted our experiments with subsets of the data which vary in size. In particular, for each type, we run 10 experiments where the subset of the training data uses n percent of the original training data set and n is in the range of [10, 100]. 10-fold cross-validation

is performed for each n.

Figure 3.4 presents the results for $KnowMore_{match}$, where the X-axis indicates the size of the training data and the Y-axis shows the P/R/F1 scores. The F1 score reaches 0.8 (0.88) at 60% (40%) percent of training data, and the P, R and F1 curves become steady with training data sets of the size of 80%(50%) for the *Movie* (*Book*) type, i.e. training data sets of at least 130 (119) entity descriptions for the *Movie* (*Book*) type.

Similar characteristics can be observed for the $KnowMore_{class}$ approach in Figure 3.5, where the F1 score reaches 0.89 (0.87) at 70% (70%) of training data for *Movie (Book)*. Hence, results suggest that even with comparably limited amounts of training data, reasonable performance can be achieved, thereby supporting the application across types and datasets. For instance, the cost for retrieving 80% (70%) of our entity matching (data fusion) ground truth from CrowdFlower with the approach as described in Section 3.5.2, is less than 15 USD and the time required is less than 24 hours for each type.

3.7.2 Model Performance across Types

In this section, we assess the performance of $KnowMore$ across different types, i.e. without a type-specific training phase. Thus, we merge the aforementioned type-specific datasets *Book* and *Movie* and perform a 10-fold cross-validation using the query sets for both types. The averaged performance of $KnowMore_{match}$ on this cross type dataset is P = 0.782, R = 0.892, F1 = 0.833, where the precision is lower than the type-specific results (Section 3.6), but the overall performance and F1 score is still comparable, the latter being slightly above the F1 score of the type-specifically trained *Movie* model (0.83). The result of $KnowMore_{class}$ is P = 0.902, R = 0.825, F1 = 0.862. Here, the precision is higher than the type-specific result for *Book* model (0.886) and lower than the one from the *Movie* model (0.967). This suggests that our approach can work on models trained on cross-type data. In order to fully validate this finding, further studies are required with more diverse query sets as well as datasets involving larger amounts of types.

3.8 Discussion & Limitations

3.8.1 Potential of KBA from Web Markup

Beside the specific KBA task evaluated in this work where we aim at (i) augmenting existing entities in a given KB by (ii) populating a given set of properties from a given KB schema for these entities, markup data shows large potential to augment KBs with properties and entities not yet present in KBs. Investigating the data from our two datasets (Movie, Book) and another set of 30 randomly selected entities of

type *Product*, we observe that a large proportion of statements in the WDC dataset involve properties not yet present in any of the KB schemas. For instance, for movies (books), 62.5% (66.8%) of entity descriptions in F contain facts not yet present in our set of mapped predicates. Comparing product descriptions from F, we detect 20.6% statements containing properties not yet present in the DBpedia ontology at all (verified through manual inspection).

In order to better highlight the potential of Web markup data to support knowledge base augmentation, we apply $KnowMore$ on all the movie and book entities from DBpedia. Out of all the 106,613 movie entities in DBpedia (10 Jan., 2017 version), we found coreferences in our markup corpus for 101,069 distinct entities (94.8%). Out of 35,577 book entities in DBpedia, we found coreferences for 34,964 entities in our markup corpus M (98.3%). In total, this resulted in 4,412,337 (1,783,231) instances, i.e. markup nodes, and 42,624,281 (16,580,862) candidate facts for the selected set of movies (books). On average, we found 5.06 (7.98) facts for each movie (book) instance. Based on the experimental results, our KnowMore approach obtained 511,409 (279,013) new facts for all DBpedia movie (book) entities. Note that this includes only entities already present in DBpedia. Whereas there is a wide variety of instances of both types which are not present in DBpedia but in markup, our approach could be used to populate DBpedia, in particular with less popular and long-tail entities. Thus, we observe a considerable potential for augmentation of KBs with new entities, as opposed to augmenting existing ones.

To assess performance in such cases, we randomly select 30 names of products under the requirement that each appears in at least 20 different PLDs in WDC, to ensure that there is sufficient consensus on the name being a legitimate product title. Manual inspection confirmed that none of such randomly selected products is represented in DBpedia.

By running the $KnowMore_{class}$ approach on 30 selected product entities, we found 136 correct facts in total, resulting in 4.53 facts for each entity on average. Table 3.12 shows the performance of $KnowMore_{class}$ and our baselines on this dataset.

Table 3.12 Data fusion performance for *Product* entities.

Approach	P	R	F1
KnowMore$_{\text{class}}$	**0.983**	**0.927**	**0.954**
PrecRecCorr	0.827	0.485	0.611
CBFS	0.876	0.686	0.769

Results indicate that the performance gain of our approach is particularly evident on such long-tail entities as represented in our *Product* dataset.

3.8.2 Limitations

Results demonstrate that *KnowMore* is able to exploit Web markup data for KBA tasks. Further improvement can be gained by applying our approach on a focused crawl, targeted towards a specific KBA task, such as movie enrichment, rather than a cross-domain Web crawl such as the WDC/Common Crawl.

In contrast to related KBA approaches such as [KM12] or [WGM+14], it is worth noting that our approach is trained for particular entity types only, not towards particular properties, as is the case with the aforementioned approaches. Hence, *KnowMore* can be adapted to a wider range of scenarios with less effort than previous KBA approaches. In addition, we have demonstrated in Section 3.7.2 that our models can potentially generalise across types.

Performance strongly differs between query sets, and hence, type-specific markup datasets, what presumably is caused by the variance in quality and quantity of facts in the WDC corpus between distinct types. Particular challenges arise from entities with a large amount of coreferences, where data usually originates from a wide variety of sources with varying degrees of quality. Compared to the baselines, our results indicate a particular strong performance gain of our approach in such cases.

Another limitation is our exclusive focus on *schema.org* statements. This constraint is motivated by the costliness of providing high-quality schema mappings between markup statements and three KBs and the fact that *schema.org* is the vocabulary of most widespread use [MRP16]. While *schema.org* adopters usually are motivated by the goal to improve their search result rankings, one assumption is that other vocabularies might show a different distribution of types and predicates, due to distinct motivations. This deserves deeper investigation as part of future work.

In this context, it is worth noting that our KBA task setup ignored a large part of the markup data, i.e. 49.3% of facts in our type-specific subsets do not involve any of our selected *schema.org* properties. To consider other vocabularies, we are currently aiming at including a preliminary schema matching step with the intention of improving recall further.

Another important aspect concerns the temporal nature of fact correctness, specifically for highly dynamic predicates, such as the price tag of a particular product. While we do not consider temporal features as such, we argue that the dynamic nature of markup annotations is well-suited to augment particularly dynamic statements. This suggests particular opportunities for updating or complementing KBs with dynamic knowledge sourced from Web markup.

4

Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web

Section 1.1 explains the motivation behind our work on improving knowledge assessibility on the Web for human. We focus specifically on the scenario of *search as learning*, as Web search became one of the most frequently used applications in everyday life. In this chapter, we aim at extending the understanding of human learning in Web search through in-depth analysis of user behavior in intentional learning-related search sessions.

Research Questions and Original Contributions. This work aims at filling this gap by contributing novel insights on the nature of knowledge gain in informational search sessions on the web, and the corresponding behavior of users. In order to carry out the analysis, we collected a dataset that simulates a real-world information search process through crowdsourcing. By combining qualitative and quantitative analysis, we seek to answer the following research questions.

RQ1: How does a user's knowledge evolve through the course of an informational search session on the web?

To further the current understanding of the impact of informational search on a user's knowledge, we recruited 500 distinct users from a crowdsourcing platform and orchestrated search sessions spanning 10 different information needs. By employing scientifically formulated *knowledge tests* to calibrate a user's knowledge before a search session, and assess it after the session, we were able to quantify knowledge gain. We found that nearly 70% of the users exhibited a knowledge gain at the end of a search session corresponding to an information need, with an overall average knowledge gain of almost 20%.

RQ2: How does the topic and information need in a search session influence a user's knowledge gain?

We explored the impact of information need on the knowledge gain of users. Our findings revealed that the information need does not directly effect the knowledge gain of users. However, we found a strong negative linear relationship between the knowledge gain of users in an informational search session and their topic familiarity. This suggests that users exhibited a higher knowledge gain in search sessions corresponding to information needs that they were less familiar with.

RQ3: What is the impact of information need on the search behavior of users in a search session?

We analyzed the search behavior of users and found a significant effect of the information need on the number of queries entered by users, the number of unique terms in their queries, the number of web pages that users navigated to, and the distinct pay-level domains accessed. Information need also had a significant effect on the amount of time users actively spent on the search results page. We also found that on average the last queries entered by users were significantly longer than the first queries across all information needs, suggesting an impact of the information consumed through the course of a search session.

4.1 Related Works

Some previous works have focused on studying the correlation between learning progress and user activity features and resource features.

Bhattacharya et al. [BG19] investigated the relationship between users' search and eye gaze behaviours and their learning performance based on a lab study (n=30). Eickhoff et al. [ETWD14] investigated the correlation between a number of features of the search session as well as the Search Engine Result Pages (SERPs) with learning needs related to either procedural or declarative knowledge. Results obtained from an analysis of large-scale query logs showed the distinct evolution of particular features throughout search sessions and the correlation of document features with the actual learning intent. The influence of distinct query types on knowledge gain was studied by Collins-Thompson et al. [CTRHS16], finding that intrinsically diverse queries lead to increased knowledge gain.

Studies on exploratory search have also investigated a similar set of search behaviors that influence the learning outcome. Hagen et al. [HPV+16] investigated the relation between the writing behavior and the exploratory search pattern of writers. The authors revealed that query terms can be learned while searching and reading.

Vakkari [Vak16] provided a structured survey of features indicating learning needs as well as user knowledge and knowledge gain throughout the search process. By matching the learning tasks into different learning stages of Anderson and Krathwohl's taxonomy [AKA+01], Jansen et al. studied the correlation between search behaviors of 72 participants and their learning stage [JBS09]. They showed that information searching is a learning process with unique searching characteristics corresponding to particular learning levels.

White et al. [WDT09] investigated the difference between the behavior of domain experts and non-experts in seeking information on the same topic. By analyzing the activity log of experts and non-experts across different domains, the authors found that the distribution of features such as number of queries and query length differed across the levels of expertise.

Gwizdka and Spence [GS06] study the behavior of users when dealing with tasks of different self-assessed difficulty and of different objective complexity. The investigation is based on a lab-study with 27 undergraduate psychology students. They showed that a searcher's perception of task difficulty is a subjective factor that depends on the domain knowledge and some other individual traits.

The aforementioned prior works consider a limited set of features or address specific learning scenarios and learning types. In our work we use a dataset that simulates a real-world information search process and present an analysis of the relation between a large number of user interaction features and quantifiable knowledge gain across topics.

4.2 Obtaining Search Session Data

We adopted a crowdsourcing approach and orchestrated search sessions with varying information needs. All interactions of the users during the search sessions were logged. We analyzed the data to further the understanding of user knowledge evolution in informational search sessions on the Web. In this section, we describe the study design and experimental setup.

4.2.1 Study Design

We recruited participants from CrowdFlower[1], a premier crowdsourcing platform. At the onset, workers were informed that the task entailed 'searching the Web for some information'. Workers willing to participate were redirected to our external platform, *SearchWell*. Figure 4.1 presents the workflow of participants in the experimental setup orchestrating informational search sessions, which consists of 5 steps: (1) Workers are recruited from the CrowdFlower platform, and those willing to par-

[1]Formerly http://www.crowdflower.com/, now https://www.figure-eight.com/

ticipate are redirected to *SearchWell*. (2) Participants are asked to answer a few questions (*knowledge test*) regarding a topic; this is used to calibrate their knowledge before the search session. (3) Participants indulge in an informational search session to satisfy a well-defined information need. (4) Participants are asked to complete a post-session test that is identical to the calibration test. (5) Participants receive a completion code, which they enter on CrowdFlower to claim their reward.

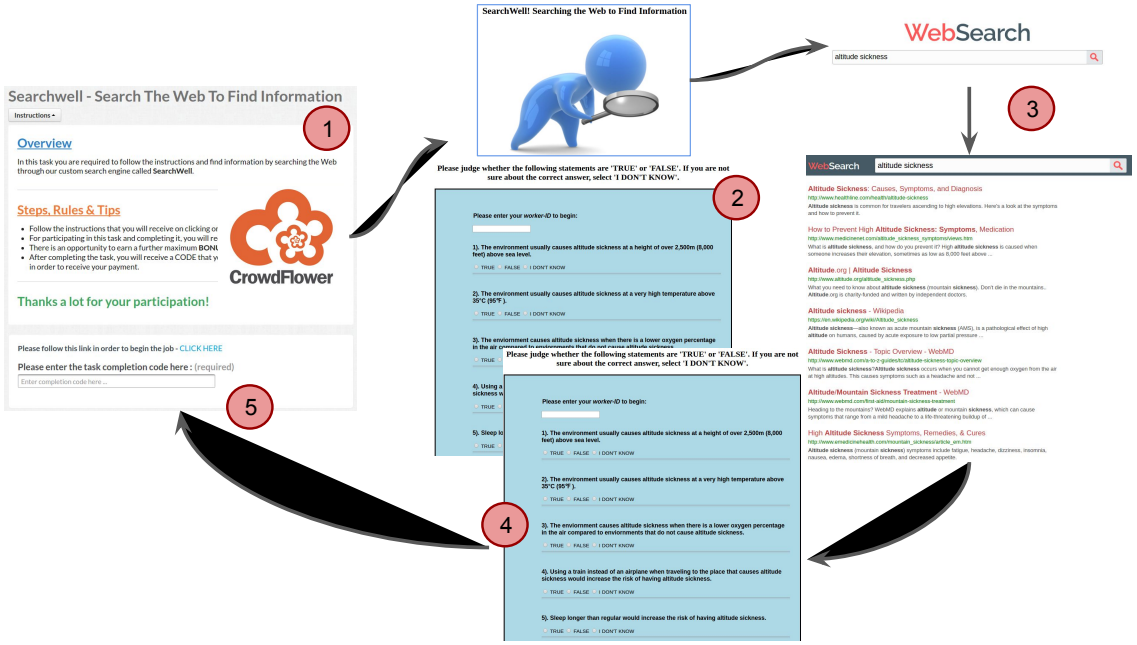


Figure 4.1 Workflow of participants in the experimental setup orchestrating informational search sessions.

Workers were first asked to respond to a few questions (technically referred to as '*items*') corresponding to a particular topic without searching the Web for answers. The questions took the form of statements pertaining to a topic, and workers had to select whether the statement was 'TRUE', 'FALSE', or 'I DON'T KNOW' in case they were not sure. In this way, we calibrated the knowledge of users corresponding to a given topic. To encourage the workers to respond without external consultation, we informed them that their responses to these questions would not affect their pay. We also encouraged workers to provide responses to the best of their knowledge and avoid guessing. The results of this pre-test were used to calibrate the knowledge of the workers with respect to the topic. We describe the topics and how the knowledge tests were created in the following Section 4.2.2. On completing the knowledge calibration test, workers were presented with their actual task.

Workers were presented an *information need* corresponding to the topic of the calibration test they completed. They were told to use the *SearchWell* platform to search the Web and satisfy their information need. To incentivize workers towards realistic attempts to learn about the topic, we informed them that they will have

to complete a final test on the topic to successfully finish the task. Furthermore, workers were conveyed the message that depending on their accuracy on the final test they could earn a bonus payment. We subsequently logged all the activities of the workers (mouse movements, key presses and clicks) within the *SearchWell* platform. Workers were allowed to begin the final test anytime after a search session, which is when a link to the final test was made available. Workers were encouraged to proceed to the next stage only once they felt that their information need was satisfied and when they were ready for the post-session test. On completing the post-session test, workers received a unique code that they could enter on CrowdFlower to claim their reward.

We restricted the participation to workers from English-speaking countries to ensure that they understood the task and instructions adequately [GYB17]. To ensure reliability of the resulting data, we restricted the participation to *Level-3 workers*[2] on CrowdFlower.

4.2.2 Topics – Defining Information Needs

We constructed a corpus of topics representing varying scopes of information needs (with some relatively broader than others). Topics were selected randomly from the *TREC 2014 Web Track* dataset[3], and corresponding information needs were defined accordingly. In all cases, the knowledge of users before beginning an informational search session was assessed using pre-tested and evaluated *knowledge tests*. Knowledge tests are scientifically formulated tests that measure the knowledge of a participant on a given topic (for example, the HIV knowledge test [CMBJ97]).

Knowledge on all given topics was measured using knowledge tests comprising of between 10 and 20 items. The answer options were in all cases 'TRUE', 'FALSE', and 'I DON'T KNOW'. The differences in the number of items reflects our attempt to feature varying scopes of information needs; relatively narrow (e.g., *Carpenter Bees*–10 items) as well as broad (e.g., *NASA Interplanetary Missions*–20 items). In the construction of all scales, an item pool comprising of more items than finally used was constructed. After a pilot test with 100 distinct participants recruited via CrowdFlower for each of the 10 topics, items that proved to be either too easy (e.g., more than 80% correct answers) or too hard/ambiguous (e.g., more false than true answers) were discarded. Table 4.1 presents the topics and corresponding information needs considered for orchestrating the informational search sessions. It also shows the internal reliability (using Cronbach's α) of the pre- and post-session knowledge tests corresponding to each topic. We observe moderate to high values of α in the pre- and post session knowledge tests, suggesting a desirable level of internal consistency.

[2] *Level-3 contributors* on CrowdFlower comprise workers who completed over 100 test questions across hundreds of different types of tasks, and have a near perfect overall accuracy. They are workers of the highest quality on CrowdFlower.

[3] http://www.trec.nist.gov/act_part/tracks/web/web2014.topics.txt

Table 4.1 Topics and corresponding information needs presented to participants in the informational search sessions, along with the internal reliability of the corresponding knowledge tests. 'α1', 'α2' represent Cronbach's α for the pre-session test and post-session test respectively. 'N' is the number of reliable participants after filtering.

Topic	Information Need	α1	α2	N
1. Altitude Sickness	In this task you are required to acquire knowledge about the symptoms, causes and prevention of altitude sickness. (20 items)	0.59	0.79	47
2. American Revolutionary War	In this task, you are required to acquire knowledge about the 'American Revolutionary War'. (10 items)	0.74	0.55	42
3. Carpenter Bees	In this task, you are required to acquire knowledge about the biological species 'carpenter bees'. How do they look? How do they live? (10 items)	0.79	0.58	46
4. Evolution	In this task, you are required to acquire knowledge about the theory of evolution. (12 items)	0.55	0.72	45
5. NASA Interplanetary Missions	In this task, you are required to acquire knowledge about the past, present, and possible future of interplanetary missions that are planned by the NASA. (20 items)	0.80	0.75	42
6. Orcas Island	In this task you are required to acquire knowledge about the Orcas Island. (20 items)	0.91	0.85	39
7. Sangre de Cristo Mountains	In this task, you are required to acquire knowledge about 'Sangre de Cristo' mountain range. (10 items)	0.70	0.52	40
8. Sun Tzu	In this task, you are required to acquire knowledge about the Chinese author Sun Tzu - about his life, his writings, and his influence to the present day. (15 items)	0.81	0.63	37
9. Tornado	In this task, you are required to acquire knowledge about the weather phenomenon that is called 'tornado' (20 items)	0.82	0.62	40
10. USS Cole Bombing	In this task, you are required to acquire knowledge about the 2000 terrorist attack that came to be known as the 'USS Cole bombing'. (10 items)	0.83	0.55	42

4.2.3 Search Environment and Data Collection

We built *SearchWell* on top of the Bing Web Search API. We logged user activity on the platform including mouse movements, clicks, and key presses, using PHP/-Javascript and the jQuery library.

To further ensure the reliability of responses and the behavioral data thus produced in the search sessions, we filtered workers using the following criteria.

- Workers who entered no queries in the *SearchWell* system. Since the aim of our work is to further the understanding of how the knowledge state of a user evolves in informational search sessions, we discard those users who did not enter a search query.
- Workers who selected the same option; either 'YES' or 'NO', for all items in the knowledge calibration test or the post-session test.
- Workers who did not complete the post-session test.

We filtered out 80 workers due to the aforementioned criteria, resulting in 420

workers across the 10 topics. The analysis and results presented hereafter are based on these 420 workers alone. For the benefit of further research in this community, the filtered data has been thoroughly anonymized and made publicly available[4]. We henceforth refer to these filtered workers as users in our experimentally orchestrated information search sessions.

4.3 Understanding Knowledge Gain

4.3.1 Measuring Knowledge Gain

We measure the knowledge gain of users in search sessions corresponding to a given information need as the difference between their knowledge calibration score and the post-session test score[5]. Table 4.2 presents the average knowledge calibration scores, post-session test scores, and the resulting knowledge gain of users across the search sessions corresponding to different information needs. Across all topics and search sessions, we found that users exhibited an average knowledge gain of nearly 20%. Nearly 70% of all the workers exhibited a knowledge gain, while the remaining workers did not. The standard deviation observed in the knowledge gain of users across all topics is notably high, due to the varying domain knowledge of users. This is evident from the average calibration scores in Table 4.2. We found that on average, the highest knowledge gain was observed through the search sessions corresponding to the topic, '*Orcas Island*', while the least knowledge gain was observed through those corresponding to the topic, '*Evolution*'. These findings are explored further in the next section.

4.3.2 Topic Familiarity vs. Knowledge Gain

We intuitively reason that users have varying levels of knowledge about a given topic, and their familiarity with the topic influences their behavior in an informational search session [LB08, GS06]. The accuracy of users in a knowledge test corresponding to a topic would therefore be a reflection of their domain knowledge. We build on this notion, and investigate the relationship between the average knowledge gain of users through informational search sessions and their average *topic familiarity*. We compute familiarity scores for different topics, by using the accuracy of users during the creation of knowledge tests. In addition, we argue that the more familiar a topic appears to be, the less prone users are to selecting the 'I DON'T KNOW' option. Table 4.3 presents the average familiarity scores and the fraction of 'I DON'T KNOW' responses ($\%IDK$) corresponding to each topic considered in our experimental setup.

[4] https://sites.google.com/view/knowledge-gain

[5] We consider the 'I DON'T KNOW' options that were selected, as incorrect responses while computing the knowledge calibration scores and post-session test scores.

Table 4.2 The average knowledge gain of users across the different topics. To enhance readability, the rows have been ordered by ascending knowledge gain *(KG)*.

Topic / Information Need	**Avg. Calibration Score** *(in %)*	**Avg. Post Score** *(in %)*	**Knowledge Gain** *(in %)*
Evolution *(N=45)*	34.07 ± 17.99	48.15 ± 22.49	14.07 ± 18.66
NASA Interplanetary Missions *(N=42)*	38.1 ± 20.53	52.5 ± 17.43	14.40 ± 22.10
Altitude Sickness *(N=47)*	55.88 ± 16.31	70.66 ± 19.11	14.78 ± 17.76
Sangre de Cristo Mountains *(N=40)*	33.25 ± 22.40	49.75 ± 18.10	16.50 ± 22.31
Tornados *(N=40)*	34.44 ± 21.02	53.47 ± 16.28	19.03 ± 22.01
Sun Tzu *(N=37)*	40.54 ± 23.37	60.18 ± 17.15	19.64 ± 21.59
American Revolutionary War *(N=42)*	34.52 ± 25.65	55.95 ± 20.71	21.43 ± 27.31
Carpenter Bees *(N=46)*	45.65 ± 27.08	67.17 ± 20.29	21.52 ± 30.50
USS Cole Bombing *(N=42)*	30.95 ± 25.22	54.37 ± 16.29	23.41 ± 31.30
Orcas Island *(N=39)*	34.74 ± 30.08	65.51 ± 22.04	30.77 ± 30.25
Overall *(N=420)*	38.22 ± 22.96	57.77 ± 18.99	19.56 ± 24.38

We thereby investigated the relationship between knowledge gain and topic familiarity, along with the %IDK. Using Pearson's correlation coefficient, we found a strongly negative linear relationship between the knowledge gain of users in informational search sessions and the topic familiarity; $\mathbf{R}= -0.87, R^2 = 0.78, p < 0.001$. This suggests that the more popular a topic is, or the more familiar that users are with a topic, the lesser they tend to learn about the topic in informational search sessions. Thus, we found that 78% of the variance in the knowledge gain of users can be explained by the topic familiarity. This is further corroborated by the moderately positive linear relationship we found between the knowledge gain of users and the fraction of *IDK* responses in the knowledge test; $\mathbf{R}= 0.72, R^2 = 0.54, p < 0.05$. An intuitive explanation for this observation is that the lesser a user knows about a topic, the more there is to learn through an informational search session, increasing the scope for knowledge to be gained.

We conducted a one-way between users ANOVA to investigate the effect of topics on the knowledge gain of users. Our findings revealed a lack of significant effect of topics on the knowledge gain of users across the 10 topic conditions.

4.3.3 User Queries and Click Behavior

User Queries. We found that on average users collectively fired 92 distinct queries across the different topics, and corresponding to the different information needs in each search session. Table 4.4 presents our findings pertaining to user queries. We note that on average users entered at least 2 distinct queries in a search session, with

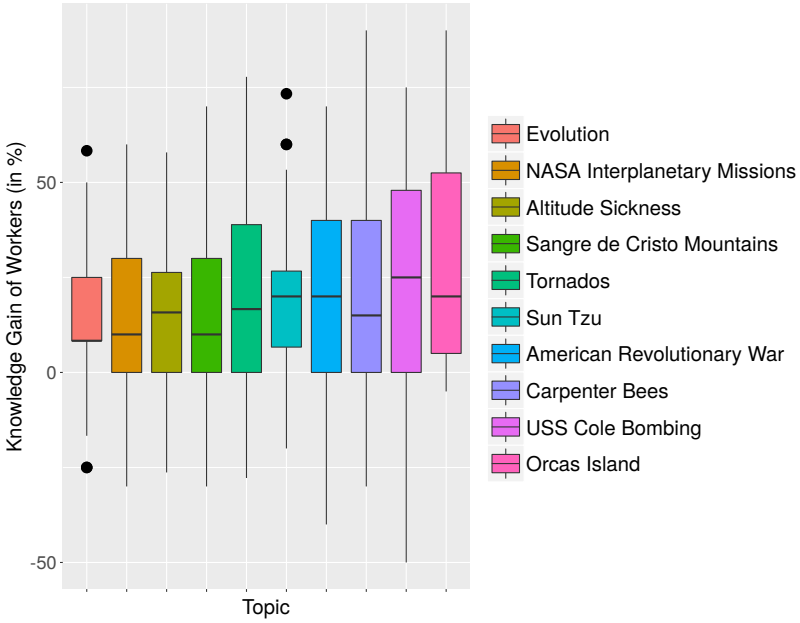


Figure 4.2 The average knowledge gain of users across the different topics (in ascending order of knowledge gain).

an average query length of just over 4 terms. Users employed a minimum of 6 unique terms in their queries on average. We conducted a one-way ANOVA to investigate the effect of topic on the number of queries fired by users. We found a significant difference in the number of queries entered by users across the 10 topical conditions at the $p < 0.001$ level; $F(9, 419) = 3.941$. Post-hoc comparisons using the Tukey-HSD test revealed significant differences in the number of queries entered by users at the $p < 0.001$ level corresponding to the topic '*NASA Interplanetary Missions*' in comparison to each of '*Altitude Sickness*', '*American Revolutionary War*', '*Carpenter Bees*' and '*USS Cole Bombing*'. Similarly, significant differences were revealed at the $p < 0.05$ level corresponding to the topic '*NASA Interplanetary Missions*' in comparison to '*Sangre de Cristo Mountains*', and '*American Revolutionary War*' in comparison to '*Evolution*'. On investigating the relationship between the number of queries entered by users and the knowledge gain through a search session, we did not find any significant linear relationship using Pearson's **R**. This suggests that although the information need in a search session influences the number of queries entered by users, there is no measurable effect of the number of queries on the knowledge gain of users.

To analyze the effect of the topics on the number of unique query terms entered by users, we conducted a one-way between users ANOVA. We found a significant difference in the number of unique terms used by users across the 10 topic conditions at the $p < 0.001$ level; $F(9, 419) = 5.44$. Post-hoc comparisons using the Tukey-HSD test revealed a significant difference in the number of unique terms entered by users

Table 4.3 The average knowledge gain of users in comparison to the topic familiarity, and the percentage of 'I DON'T KNOW' (IDK) responses to questions in the knowledge test.

Topic/Information Need	Knowledge Gain	Familiarity	% IDK
Altitude Sickness	14.78	52.49	28.67
American Revolutionary War	21.43	40.77	43.57
Carpenter Bees	21.52	42.18	33.57
Evolution	14.07	48.28	31.41
NASA Interplanetary Missions	14.40	51.91	38.61
Orcas Island	30.77	34.69	50.17
Sangre de Cristo Mountains	16.50	44.41	42.15
Sun Tzu	19.64	50.38	40.00
Tornados	19.03	46.2	33.65
USS Cole Bombing	23.41	39.93	48.47

Table 4.4 Queries fired by users in informational search sessions corresponding to different topics. Note that the query length is measured in 'terms'. For readability, the rows have been ordered by an increasing knowledge gain *(KG)*. In the heading, *DQ* refers to distinct queries and *UT* refers to unique terms.

Topic/Information Need	KG *(in %)*	#DQ	DQ Per User	Query Length	#UT	UT Per User
Evolution	14.07	140	3.11 ± 2.93	5.62 ± 2.69	437	4.70 ± 1.89
NASA Interplanetary Missions	14.40	160	3.81 ± 4.17	3.58 ± 1.55	671	4.33 ± 3.60
Altitude Sickness	14.78	80	1.70 ± 1.24	4.29 ± 2.56	221	3.57 ± 1.88
Sangre de Cristo Mountains	16.50	72	1.80 ± 1.49	4.31 ± 2.70	238	4.21 ± 2.78
Tornados	19.03	90	2.25 ± 3.68	4.41 ± 2.59	150	9.71 ± 14.81
Sun Tzu	19.64	85	2.30 ± 2.70	8.15 ± 5.52	320	15.98 ± 25.75
American Revolutionary War	21.43	59	1.40 ± 0.76	2.36 ± 0.91	182	3.69 ± 3.73
Carpenter Bees	21.52	75	1.63 ± 1.05	4.74 ± 2.15	164	5.95 ± 5.31
USS Cole Bombing	23.41	71	1.69 ± 1.03	6.25 ± 3.80	177	8.65 ± 11.56
Orcas Island	30.77	91	2.33 ± 2.80	1.93 ± 1.84	144	3.75 ± 10.74
Overall	19.56	92.30	2.20 ± 2.18	4.56 ± 2.63	270.40	6.45 ± 8.20

corresponding to the topic, '*NASA Interplanetary Missions*' in comparison to each of the other topics except '*Evolution*' and '*Sun Tzu*' at the $p < 0.001$ level. We did not find a significant linear relationship between the number of unique query terms entered by users and their knowledge gain in search sessions, using Pearson's **R**.

Evolution of Query Terms and Lengths. We investigated how queries from users evolved within a search session corresponding to a given information need. Table 4.5 presents our findings, considering only those users who entered 2 or more queries in a particular search session. We note that on average across all the topics, the last query entered by users is longer (5.11 terms) than the first query (4.53 terms) entered in the session.

We also note that on average, the number of unique terms in the last query is greater than the number of unique terms entered by users in their first query. A two-

Table 4.5 A comparison of the first and last query lengths (QL), number of unique terms (UT) in the first and last query entered by users within search sessions.

Topic/Information Need	First QL	Last QL	First UT	Last UT
Altitude Sickness	4.25 ± 2.38	4.38 ± 2.32	3.50 ± 1.58	3.56 ± 1.62
American Revolutionary War	4.73 ± 2.05	5.18 ± 2.37	4.09 ± 1.38	4.55 ± 1.97
Carpenter Bees	5.12 ± 3.29	5.29 ± 2.84	3.59 ± 1.82	3.65 ± 1.64
Evolution	3.00 ± 1.80	5.65 ± 4.75	2.46 ± 1.87	4.15 ± 3.17
NASA Interplanetary Missions	6.96 ± 5.96	6.92 ± 5.16	4.24 ± 3.20	4.80 ± 3.05
Orcas Island	2.58 ± 1.26	3.08 ± 2.06	2.58 ± 1.26	2.83 ± 1.40
Sangre de Cristo Mountains	4.15 ± 1.03	4.08 ± 1.44	4.15 ± 1.03	3.92 ± 1.27
Sun Tzu	6.88 ± 5.86	7.75 ± 6.77	4.31 ± 2.54	4.69 ± 2.89
Tornados	2.30 ± 1.68	3.60 ± 2.33	2.10 ± 1.30	2.90 ± 1.30
USS Cole Bombing	5.29 ± 4.16	5.18 ± 4.15	4.35 ± 3.32	4.41 ± 3.34
Overall	4.53 ± 2.95	5.11 ± 3.42	3.54 ± 1.93	3.95 ± 2.17

tailed T-test revealed that this difference is statistically significant; $t(413) = 3.99, p < 0.05$. This suggests that as the users consume more information through the course of a search session related to a given topic, their queries tend to become longer.

User Clicks. We analyzed the clicks of users on results corresponding to each of the queries they entered within search sessions. Table 4.6 presents our findings with respect to the average number of clicks per user, clicks per query, the average rank of the results that were clicked, and the average interval of time between two consecutive clicks on search results. We note that users clicked on just over 2 search results on average, and on at least 1 result per query on average. In line with prior works that analyzed user behavior with search results [GJG04, ABDR06], we found that users in the informational search sessions orchestrated in our experiments, typically clicked on top-ranked results (with an average rank of 2.18). The average interval between two clicks by a user in a search session was found to be 0.69 minutes. On investigating the linear relationship between the click interval and knowledge gain of users, we found no significant correlation.

4.3.4 Session Duration and Browsing Behavior

Session Length. We analyzed the session lengths[6] of users and their browsing behavior in informational search sessions corresponding to the different topics. Our findings are summarized in Tables 4.7 and 4.8. We found that the average session

[6]For a given topic and user, we measured the session length as the time from which the first query was entered in *SearchWell* by the user after the calibration test, until the time at which the last webpage accessed by the user was active before the post-session test. Note that users were allowed to carry out only one search session.

Table 4.6 The average number of clicks per user, clicks per query, the average rank of the results clicked by users, and the average interval between two consecutive clicks on the search results (in mins) across different topics.

Topic / Information Need	#Clicks Per User	#Clicks Per Query	Rank of Result Clicked	Click Interval
Altitude Sickness	2.49 ± 1.38	1.88 ± 1.36	2.96 ± 2.30	1.32 ± 1.84
American Revolutionary War	2.05 ± 1.59	1.64 ± 1.41	1.84 ± 2.15	0.94 ± 1.58
Carpenter Bees	1.80 ± 1.44	1.32 ± 1.09	1.78 ± 1.53	0.49 ± 0.94
Evolution	3.36 ± 2.84	1.61 ± 1.45	2.66 ± 2.04	0.33 ± 0.69
NASA Interplanetary Missions	2.90 ± 2.09	1.44 ± 1.38	2.64 ± 2.40	0.49 ± 0.72
Orcas Island	2.03 ± 1.94	1.35 ± 1.08	3.17 ± 2.17	0.58 ± 1.01
Sangre de Cristo Mountains	2.38 ± 2.23	1.76 ± 2.11	1.66 ± 1.14	1.14 ± 1.97
Sun Tzu	2.19 ± 1.74	1.37 ± 0.92	1.90 ± 1.28	0.64 ± 1.09
Tornados	1.83 ± 1.50	1.43 ± 1.42	1.65 ± 1.11	0.79 ± 1.28
USS Cole Bombing	1.81 ± 1.47	1.37 ± 1.23	1.52 ± 1.04	0.20 ± 0.50
Overall	2.28 ± 1.82	1.52 ± 1.34	2.18 ± 1.72	0.69 ± 1.16

length of users across the different topics was nearly 5 mins long. To understand the effect of the 10 topics considered on the session length exhibited by users, we conducted a one-way ANOVA. Results revealed no significant effect of the topics on the session length. We also analyzed the relationship between the session length of users and the knowledge gain using Pearson's **R**. We did not find a significant linear relationship between these variables, suggesting that length of a session does not directly influence the knowledge gain of users.

Table 4.7 The average session lengths (*SL*) of users across different topics, the session length per query, the number of webpages navigated to from the results page (*#Pages Navigated*), the number of webpages navigated to per query entered, and the active time spent on a webpage.

Topic / Information Need	Session Length (*SL*) (in mins)	SL Per Query (in mins)	#Pages Navigated	#Pages Per Query	Active Time Per Page (in mins)
Altitude Sickness	4.75 ± 3.48	3.45 ± 2.91	4.74 ± 2.18	3.46 ± 1.93	2.27 ± 1.57
American Revolutionary War	3.88 ± 2.86	3.02 ± 2.22	4.31 ± 1.64	3.36 ± 1.10	2.06 ± 1.37
Carpenter Bees	3.30 ± 2.67	2.19 ± 1.58	4.57 ± 2.13	3.20 ± 1.39	1.72 ± 1.42
USS Cole Bombing	3.96 ± 3.62	2.88 ± 2.98	4.31 ± 2.02	3.03 ± 1.40	1.72 ± 1.24
Evolution	6.79 ± 9.28	2.55 ± 3.12	7.04 ± 5.56	2.87 ± 1.38	1.86 ± 1.29
NASA Interplanetary Missions	6.64 ± 5.68	2.41 ± 2.89	7.79 ± 5.62	2.92 ± 1.79	2.22 ± 1.79
Orcas Island	6.52 ± 12.79	3.30 ± 3.14	5.67 ± 4.66	3.50 ± 2.06	2.25 ± 2.08
Sangre de Cristo Mountains	4.91 ± 4.35	3.29 ± 3.16	5.78 ± 3.37	3.93 ± 2.24	1.90 ± 1.51
Sun Tzu	3.87 ± 4.11	1.99 ± 1.61	5.22 ± 3.19	2.98 ± 1.06	1.82 ± 1.59
Tornados	3.57 ± 3.20	2.64 ± 2.87	5.15 ± 3.71	3.43 ± 1.63	1.85 ± 1.30
Overall	4.82 ± 5.20	2.78 ± 2.65	5.46 ± 3.41	3.27 ± 1.60	1.97 ± 1.52

Navigation. During the search sessions corresponding to the different topics, users navigated to over 5 webpages on average (as shown in Table 4.8). To understand the effect of the 10 different topics on the navigation behavior of users, we conducted

a one-way between users ANOVA. Results confirmed a significant difference in the number of pages users navigated to across the 10 different topic conditions at the $p < 0.001$ level; $F(9, 419) = 9.154$. Post-hoc comparisons using the Tukey-HSD test revealed that the number of webpages navigated to by users in the search sessions corresponding to the topic of '*Evolution*' was significantly different in comparison to all other topics at the $p < 0.001$ level. In addition, the number of webpages navigated by users in the search sessions corresponding to the topic of '*NASA Interplanetary Missions*' was found to be significantly more than those pertaining to the topic of '*Altitude Sickness*'. We did not find a significant linear relationship between the number of webpages that users navigated to, and their knowledge gain.

Table 4.8 The average number of pay-level domains (#PLDs) accessed by users during the search session, the amount of time spent on the search engine results page (SERP), the amount of time active on the results page, and the number of pages navigated to from the results page, and other subsequent pages (non-SERPs).

Topic / Information Need	**#PLDs**	**Time Spent on SERP (in mins)**	**Time Active on SERP (in mins)**	**#Pages from SERP**	**#Pages from Non-SERPs**
Altitude Sickness	1.89 ± 1.17	9.95 ± 5.75	0.60 ± 0.44	2.11 ± 1.37	0.06 ± 0.32
American Revolutionary War	1.45 ± 0.93	7.92 ± 4.55	0.53 ± 0.44	1.62 ± 1.09	0.10 ± 0.37
Carpenter Bees	1.50 ± 1.12	7.89 ± 4.68	0.55 ± 0.51	1.52 ± 1.12	0.15 ± 0.42
Evolution	2.31 ± 1.74	10.34 ± 7.85	0.91 ± 0.65	2.67 ± 2.57	0.04 ± 0.21
NASA Interplanetary Missions	1.60 ± 1.20	10.43 ± 6.63	0.94 ± 0.97	2.17 ± 1.99	1.05 ± 3.43
Orcas Island	1.51 ± 1.32	11.08 ± 16.20	0.69 ± 0.58	1.92 ± 2.04	0.33 ± 0.69
Sangre de Cristo Mountains	1.58 ± 1.07	8.80 ± 5.00	0.51 ± 0.44	2.28 ± 1.82	0.48 ± 0.74
Sun Tzu	1.73 ± 1.39	8.82 ± 5.43	0.61 ± 0.57	1.89 ± 1.57	0.19 ± 0.46
Tornados	1.50 ± 0.89	7.89 ± 5.19	0.43 ± 0.45	1.68 ± 1.15	0.23 ± 0.61
USS Cole Bombing	1.33 ± 0.81	7.53 ± 4.22	0.46 ± 0.30	1.40 ± 0.87	0.17 ± 0.43
Overall	1.64 ± 1.16	9.06 ± 6.55	0.62 ± 0.53	1.93 ± 1.56	0.28 ± 0.77

For each query that was entered, users navigated to over 3 webpages on average. We conducted a one-way between users ANOVA to compare the effect of topics on the number of webpages navigated by users across the 10 topic conditions. We found no significant effect of such navigation behavior on the knowledge gain. We also found no significant linear relationship between these two variables using Pearson's **R**.

Next, we investigated the amount of time users actively spent on each webpage that they navigated to. We found that on average users spent almost 2 minutes per page. We compared the effect of the topics on the average amount of time that users spent on webpages across the 10 different topic conditions using a one-way between users ANOVA. We found no significant effect across the topic conditions. Using Pearson's **R**, we found a weak positive linear relationship between the amount of active time users spent on webpages and their knowledge gain; $\mathbf{R}= 0.27, R^2 = 0.07, p < 0.001$. This suggests that the amount of time that users spend actively on webpages within the search session describes around 7% of the variance in their knowledge gain.

Domains and Search Engine Results Pages. We analyzed the pay-level domains (PLDs) of search engine results pages (SERPs) consumed by users during the informational search sessions. PLDs are sub-domains of a public top-level domain, that are acquired by paying for them. PLDs typically indicate that individual user(s) or organization(s) are likely to be in control. For instance, the PLD for `www.example.com` would be `example.com`. We note that on average, users navigated to 1.64 PLDs from the search results page. To compare the effect of topics on the number of PLDs accessed by users, we conducted a one-way between users ANOVA. We found that there was a significant effect of topics on the number of PLDs accessed by users at the $p < 0.05$ level; $F(9, 419) = 9.154$. Post-hoc comparisons with the Tukey-HSD test revealed that the users navigated to more PLDs during the search sessions corresponding to the topic of '*Evolution*', when compared to topics '*American Revolutionary War*', '*Carpenter Bees*' at the $p < 0.05$ level, and '*NASA Interplanetary Missions*' at the $p < 0.01$ level. However, we did not find a significant linear relationship between the knowledge gain of users and the number of PLDs accessed by them during the informational search sessions.

Next, we investigated the amount of time that users spent on the SERPs. We differentiate between the time users actively spent exploring the snippets on the search results page, and the total amount of time spent including the idle time. We found no significant effect of the topics on the total amount of time that users spent on the search results page on average. We found that users spent almost 40 seconds actively on the search results page on average across all topics. To compare the effect of the topics on the amount of active time spent on the SERP by users, we conducted a one-way between users ANOVA. We found that there was a significant effect of topics on the amount of active time spent by users on the search results page at the $p < 0.001$ level; $F(9, 419) = 4.066$. Post-hoc comparisons with the Tukey-HSD test revealed that users spent more active time on the results page in search sessions corresponding to the topic of '*NASA Interplanetary Missions*' in comparison to that spent in the topics, '*American Revolutionary War*', '*Carpenter Bees*', '*Sangre de Cristo Mountains*', '*Tornados*', and '*USS Cole Bombing*' at the $p < 0.01$ level. Similarly, we found that users spent more active time on the SERP in search sessions corresponding to the topic of '*Evolution*' in comparison to '*Sangre de Cristo Mountains*', '*Tornados*', and '*USS Cole Bombing*' at the $p < 0.05$ level. We did not find a significant linear relationship between the knowledge gain of users and the active time spent on the results page.

We also analyzed the number of webpages that users navigated to directly from the SERP and those that users navigated to from other webpages. We found that the users navigated to nearly 2 webpages from the search results page on average across all topics, while the navigation from a non-results page was less frequent with an average of 0.28 across all topics. To compare the effect of topics on the number of pages navigated to from SERPs and non-SERPs, we conducted between users one-way ANOVAs across the 10 different topic conditions. Results confirmed a significant

effect of topics on the number of pages navigated to from the search results page in search sessions at the $p < 0.05$ level; $F(9, 419) = 2.375$. Post-hoc comparisons with the Tukey-HSD test revealed that users in search sessions corresponding to the '*Evolution*' topic navigated to significantly more pages originating from the search results page when compared to that pertaining to the topics of '*Carpenter Bees*' and *'USS Cole Bombing'*. We also found a significant effect of topics on the number of webpages users navigated to from non-SERPs at the $p < 0.01$ level; $F(9, 419) = 2.662$. Post-hoc comparisons with the Tukey-HSD test revealed that users in search sessions corresponding to the topic '*NASA Interplanetary Missions*' navigated to significantly more webpages from non-SERPs in comparison to all other topics except '*Orcas Island*' and '*Sangre de Cristo Mountains*' at the $p < 0.05$ level. This suggests that the nature of topics effects how users navigate from the search results page. Using Pearson's **R** we found no significant linear relationship between the knowledge gain of users and the number of pages navigated from either search result pages or non-SERPs.

PLDs Across Topic. We analyzed the most frequently accessed PLDs during search sessions corresponding to different topics. We found that `wikipedia.org` was the most accessed PLD, accounting for 47.5% of PLDs accessed across all topics. This was followed by `nasa.gov` (6.6%) and `healthline.com` (3.3%). The most number of distinct PLDs accessed by users corresponded to the topic of '*Evolution*', followed by '*Altitude Sickness*' and '*Sun Tzu*'.

4.3.5 Query Formulation

Query Overlap with Topic Description and Knowledge Tests. We investigated the nature of queries fired by users in the search sessions corresponding to different information needs. First, we analyzed the overlap in the query terms with the terms in the topic description, as well as the questions in the knowledge test. Since users consumed this information prior to beginning the search session, we were interested in analyzing the fraction of query terms that go beyond the terms in the topic description and knowledge tests. Our findings are presented in Table 4.9. We note that on average across all topics, almost 11% of the query terms entered by users did not overlap with the topic description or the knowledge tests. Around 55% of the query terms were present in the topic descriptions and nearly 82% overlapped with terms in the knowledge tests on average across all topics. This is understandable, considering that the pre-session calibration test also served as a guide for kindling a realistic information need among the users.

Using Pearson's **R**, we also found a positive linear relationship between the knowledge gain of users and the percentage of query terms fired by them that did not overlap with terms in either the topic description or the knowledge tests; $\mathbf{R} = 0.41, R^2 = 0.17, p < 0.001$. This suggests that the nature of the query terms that users enter in the search sessions (in terms of overlap with topic descriptions or knowledge tests)

can explain around 17% of the variance in their knowledge gain.

Table 4.9 Percentage of query terms ($\%QT$) that are distinct with respect to the terms in the topic description TD and knowledge tests KT, and the average query complexity corresponding to the different information needs.

Topic / Information Need	$\%QT$ not in TD –(1)	$\%QT$ not in KT –(2)	$\%QT$ not in (1) or (2)	Query Complexity
Altitude Sickness	9.83	19.23	4.70	19.17
American Revolutionary War	26.19	22.22	22.22	18.12
Carpenter Bees	41.44	28.18	22.65	22.00
Evolution	69.44	11.11	11.11	20.32
NASA Interplanetary Missions	71.11	22.91	5.47	19.95
Orcas Island	31.31	12.62	12.62	22.17
Sangre de Cristo Mountains	68.97	6.90	6.90	19.50
Sun Tzu	56.06	37.29	11.88	20.06
Tornados	55.81	9.30	2.33	19.18
USS Cole Bombing	16.83	10.89	9.90	19.26
Overall	44.70 ± 22.79	18.06 ± 9.69	10.98 ± 6.90	19.97 ± 1.27

Query Complexity. We also analyzed the complexity of the queries fired by users in search sessions corresponding to the different topics. We computed query complexity using the method motivated by Eickhoff et al. [ETWD14]. We rely on a listing of over 30,000 English words along with the age at which native speakers typically learn the term compiled by Kuperman et al. [KSGB12]. The higher this score, the harder and more specialized a term is assumed to be. We assume the maximum age of acquisition across all query terms as a measure of the query complexity. Table 4.9 presents the average complexity of the queries entered by users in search sessions corresponding to the different topics. Using Pearson's **R**, we found a positive linear relationship between the knowledge gain of users and their corresponding average query complexity; $\mathbf{R}= 0.50, R^2 = 0.25, p < 0.001$. This suggests that the complexity of queries entered by users during the informational search sessions in our setup, explains 25% of the variance in their knowledge gain.

Evolution of Queries Within Search Sessions. Next, we analyzed the overall evolution of queries entered by users with the search sessions across all topics. Figure 4.3a presents the number of queries entered by users across all topics corresponding to the query rank. We refer to the sequence number of the query entered by each user in a search session as the *query rank*. For example, a query rank of 5 implies the 5^{th} query that is entered in a given search session by a given user. We observe a power-law distribution, indicating many users fire only a few queries within a search session and that a few users fire many queries.

Interestingly, we found that the average complexity of the queries entered by users within the search sessions does not fluctuate significantly over time, with an average query complexity of 20.08 ± 0.38 across the query ranks. We also analyzed the evolution in the overlap of query terms with terms in the task description and

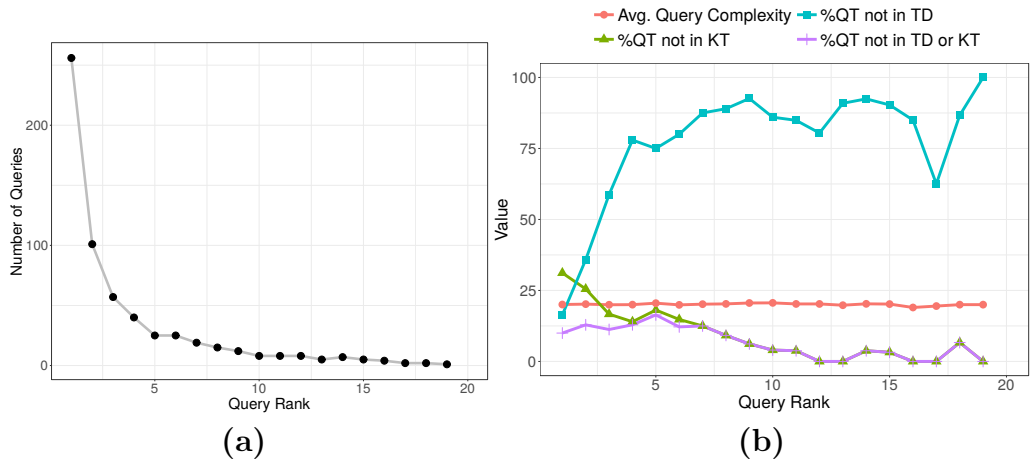


Figure 4.3 Overall evolution of queries across all topics: (a) Number of queries fired by users at a given rank across all topics within search sessions. (b) Evolution of the average query complexity, and overlap of query terms ($\%QT$) with terms in the task description (TD) and knowledge tests (KT) across all topics within the search sessions.

knowledge tests corresponding to all topics. As shown in Figure 4.3b, we found that the overlap of query terms with terms from the topic descriptions decreases with an increasing query rank. Using Pearson's **R**, we found a moderately strong positive linear relationship between the query rank and the $\%QT$ *not in* TD; $\mathbf{R}= 0.63, R^2 = 0.40, p < 0.01$. This suggests that 40% of the variance in query term overlap with the terms in the topic description can be explained by the query rank. At the same time, we found that with an increase in query rank the overlap of query terms with terms from the corresponding knowledge tests also increases. This was confirmed by a strong negative linear relationship between the query rank and $\%QT$ *not in* KT; $\mathbf{R}= -0.87, R^2 = 0.76, p < 0.001$, suggesting that over 76% of the variance in the query term overlap with terms in the knowledge tests can be explained by the query rank. The overall trend in the overlap of query terms with terms from either the corresponding task descriptions or knowledge tests TD *or* KT was similarly found to exhibit a strong negative linear relationship; $\mathbf{R}= -0.82, R^2 = 0.68, p < 0.001$.

Our findings indicate that to formulate their queries through the course of the search sessions, users on average used a decreasing number of terms from the information need presented to them, and an increasing number of terms from the pre-session calibration test they completed.

4.4 Discussion

4.4.1 Main Findings

Based on the analysis presented in previous sections, we summarise the main findings of this work as follows:

- Through our experimental results, we found that users depicted a higher knowledge gain in informational search sessions corresponding to those topics that are generally less popular[7], resulting in users having a relatively less overall *topic familiarity* with the information need. We intuitively reason that the more a user already knows about a given topic, the less he/she tends to learn through a search session on the Web.

- We found evidence which affirms that the information need in a search session influences the number of queries entered, the number of pages consumed, and the number of different PLDs accessed by users. However, these factors did not affect the knowledge gain of users through the search session.

- Users navigated to more pages from a search engine result page (SERP) in comparison to non-SERPs. We also found a significant effect of topics on navigation patterns of users; users navigated to more pages from SERPs corresponding to some topics more than they did in case of others. This however, did not have an effect on their knowledge gain.

- We found that the last queries entered in search sessions are significantly longer than the first queries, with more unique terms in the last queries than in the first. This indicates that the knowledge gained through the course of the search session, allows a user to formulate such richer queries.

  We also found that the average complexity of queries entered by users in search sessions is positively correlated to their knowledge gain, such that 25% of the variance in their knowledge gain can be explained by their average query complexity. The amount of active time that users spent on webpages also correlated positively with their knowledge gain, such that the amount of time users spent actively on webpages described around 7% of the variance in their knowledge gain.

- In line with prior works that studied user interaction with search results, we found that workers typically clicked and consumed top-ranked results on the SERP (with an average rank of 2.18).

[7]This was estimated by the overall accuracy of 100 distinct responses from crowd workers during the knowledge test formulation for each topic.

4.4.2 Contributions and Limitations

We observed that during the informational search sessions, users enter queries using terms they encountered in the knowledge tests. Although the main purpose of the pre-session tests was to calibrate the knowledge of the users, we also reasoned that the items in the knowledge test could steer users towards the diverse facets of the information need and help shaping realistic search session scenarios. This was confirmed by our findings in Section 4.3.5, where we found that users tend to employ an increasing number of terms from the pre-session calibration test in their queries, through the course of a search session.

We have considered an arguably small set of topics and corresponding information needs for our experiments in this work. However, it is noticeably challenging to create reliable knowledge tests corresponding to each topic in a manner that allows us to measure the knowledge gain of users through search sessions. Nevertheless, we have gathered a substantial amount of data from various search sessions spanning 420 reliable users across 10 topics and representing diverse information needs.

We did not find any impact of the available user demographics on knowledge gain of users across the topics. To control for Type-I error inflation in our multiple comparisons, we used the Holm-Bonferroni correction for family-wise error rate (FWER) [Hol79], at the significance level of $\alpha < 0.05$.

5

Predicting User Knowledge Gain in Informational Search Sessions

The work in this chapter continues focus on improving the knowledge assessibility on the Web for human. In Chapter 4, we analyzed the correlation of users' search behavior with their domain knowledge, influencing their knowledge gain through the course of a search session. In this chapter, we extend the work on understanding user learning in search by building machine learning models to predict user knowledge state and knowledge gain using features extracted from user interaction with search engine in a session.

Although the importance of learning as an implicit element of Web search has been established, there is still only a limited understanding of the impact of search behavior on a user's knowledge state and knowledge gain. Prior work has focused on improving the learning experience and efficiency during search sessions, but the measurement of a user's knowledge gain through the course of an informational search session has not yet been addressed. This is in part due to the difficulty in accurately quantifying knowledge gain through the course of a search session. If Web search engines that are currently optimized for relevance can be re-molded to serve learning outcomes, the capability to predict knowledge gain will be a crucial step forward.

In this chapter, we aim to address the aforementioned gap by introducing a supervised model to predict a user's knowledge state and knowledge gain from user behavior features captured during the search sessions.

Original Contributions. Through our work, we make the following contributions to the current body of literature:

- **Novel feature sets.** A novel set of user behavioral features extracted from different dimensions of a search session, namely features related to the session, queries, SERP, browsing behavior and mouse movements.
- **Knowledge state/gain prediction models.** Models for predicting the user's

knowledge gain and state during real-world informational search sessions. The experimental results underline that a user's knowledge gain and knowledge state can be modeled based on a user's online interactions observable throughout the search process.

- **Feature analysis.** An analysis of the effect of user interactions (ranging from the queries entered to their browsing behavior) on their knowledge state and knowledge gain.

Implications. The capability to predict a user's knowledge state and gain through the course of an informational search session has the potential to reshape search engines to support learning outcomes as an implicit part of retrieval and ranking. This is of particular importance given that Web search already augments learning processes in a variety of informal as well as formal learning scenarios, such as classrooms, libraries and in work environments. Our contributions advance the current understanding of learning through Web search, setting important precedents for further research.

5.1 Related Works

Apart from the previous works that focus on studying the correlation between learning progress and user activity features and resource features as introduced in Section 4.1, researchers have also proposed to use features that are extracted from search activity to measure the user's knowledge state in an online learning environment.

Syed and Collins-Thompson [SCT17] proposed to optimize the learning outcome of the vocabulary learning task by selecting a set of documents that consider the keyword density and domain knowledge of the learner. Furthermore, they explored the possibility of using regression models and features extracted from user accessed document content to predict user knowledge change on vocabulary learning tasks [SCT18]. Experimental results indicate that document content features are effective for predicting user knowledge.

Gwizdka et al. [GC16] proposed to assess learning outcomes in search environments by correlating individual search behaviors with corresponding eye-tracking measures. The user search activity is collected through a lab-experiment ($n = 30$) where the participants are asked to search for pre-defined health related tasks, and the learning outcome is assessed through quiz.

Zhang et al. [ZCB11] explored using search behavior as an indicator for the domain knowledge of a user. Through a small study ($n = 35$), they identified features such as the average query length or the rank of documents consumed from the search results as being predictive.

Further, Cole et al. [CGL+13], observed that behavioral patterns provide reliable indicators about the domain knowledge of a user, even if the actual content or topics

of queries and documents are disregarded entirely.

Other works have focused on detecting task difficulty in search environments based on user activity, where the subjective assessment of task difficulty is highly correlated to the user's domain knowledge [LB08]. Arguello [Arg14] extracted features from several dimensions of user activities (e.g. query, click) and use logistic regression for the task difficulty prediction in a search environment. Data was collected through a crowdsourcing platform, and the author used search tasks created by Wu et al. [WKEA12], which contain task difficulty assessments on multiple dimensions. The feature-evaluation result shows that the most predictive features were different for whole-session vs. fist-round prediction, that mouseover features were effective for first-round prediction, and that level of interest and prior knowledge features did not improve performance.

Our work in this chapter leverages these results to derive a comprehensive feature set to build supervised models. In contrast to prior works, we aim at predicting the knowledge state of a user – avoiding the need for explicit post-search knowledge assessments.

Futhermore, our study (Chapter 4) shows that the learning intent has a strong effect on user behavior. Given that the learning intent of the user in real search environments is diverse and impossible to foresee, in order to build generalizable prediction models it is necessary that the task dependency of features is taken into consideration. In Chapter 6, we aim at improving the generalizability of the knowledge prediction models that are introduced in Chapter 5 using topic independent features extracted from both learning resource and user interactions perspective.

5.2 Problem Definition

In the context of Web search, Broder [Bro02] classified search queries according to their intent into three classes: 1) navigational, 2) informational, and 3) transactional. Herein, *informational queries* are defined as those queries where 'the intent of a user is to acquire some information assumed to be present on one or more webpages' [Bro02]. Thus, *informational queries* imply a particular learning intent; *intentional learning* is generally defined as learning that is motivated by intentions and is goal directed [Blu12], in contrast to latent or incidental learning.

Based on the constructs of intentional learning and informational queries, we arrive at the following definition:

Definition 2 *Intentional Learning-Related Search Session. An intentional learning-related search session comprises of the sequence of a user's actions, with respect to satisfying her learning intent in a Web search environment through informational queries. A user's sequence of actions begins with querying the Web, and includes browsing through the search results, click and scroll activity, navigation via hyperlinks,*

query reformulations, and so forth.

For the sake of simplicity, we henceforth refer to informational sessions, i.e. sessions with a particular learning intent, as "sessions".

In this work, from the observed user interactions in informational search sessions, we aim to predict (i) the *knowledge state* and (ii) *knowledge gain* of a user as follows.

Definition 3 *Predicting a User's Knowledge State and Gain During Search Sessions. Let s be a search session starting at time t_i and ending at time t_j aimed at satisfying a particular information need, that is, a learning intent ι of user u. Based on the user interactions during session s captured in the time period $[t_i, t_j]$, we aim to:*

(1) classify the knowledge state (KS) $k(t_j)$ of u at time point t_j with respect to a particular information need. For the sake of this work, a user's knowledge state with respect to a particular information need is defined by the user's capability to correctly respond to a set of questions about the corresponding information need. We classify a user's knowledge state into 3 classes according to her capability: low knowledge state, moderate knowledge state and high knowledge state (Section 5.3).

(2) classify the knowledge state change, i.e. the knowledge gain (KG) $\Delta k(t_i, t_j)$ of u during time period $[t_i, t_j]$ into different degrees. Similarly, a user's knowledge gain with respect to a particular information need is defined as the improvement of user capability (accuracy) to correctly respond to a set of test questions about the corresponding information need. We classify user knowledge gain into 3 classes according to the improvement of user capability: low knowledge gain, moderate knowledge gain and high knowledge gain (Section 5.3).

5.3 Knowledge State and Knowledge Gain Classes

For the analysis and the experimental evaluation of our models, we use the dataset collected from a crowdsourcing study[1]. The detailed description of the data collection process, the data cleaning cretrias and the descriptive analysis of the dataset can be found in Section 4.2.

We used a *Standard Deviation Classification* approach to obtain three classes of learners with regard to their level of knowledge. Assuming approximately normal distributions of the respective test scores (X) for the different topics, we transformed the test scores into Z-scores with a mean of 0 and a Standard Deviation (SD) of 1 (standardization). We then used statistically defined intervals (X < -0.5 SD = low;

[1]In addtion to the previous 10 topics, we collected 100 sessions for the topic 'HIV'. The description of the corresponding task is "In this task you are required to acquire knowledge about the transmission, prevention, and consequences of HIV infection.". The corresponding knowledge test contains 45 items.

Table 5.1 User groups created based on $average \pm 0.5SD$.

Task	Mean	SD	Low	Moderate	High
KG	0.193	0.231	167	179	122
KS	0.618	0.191	145	171	152

-0.5 SD < X < 0.5 SD = moderate; 0.5 SD < X = high) for the classification of the learners into roughly equal groups with low, moderate, or high knowledge. The same procedure was repeated for knowledge gain. Here as well, the empirical knowledge gain for every test was transformed into corresponding Z-scores and three roughly equal groups (low knowledge gain; moderate knowledge gain; high knowledge gain) were defined accordingly. In view of the substantial variety of different topics, we argue that such a tripartite categorization of knowledge states and knowledge gains respectively allows for the construction of robust models, which are themselves based on a large variety of features. Thus, insights from the learning tasks considered can be generalized to other similar intentional learning activities. This procedure weighs all different knowledge tests equally irrespective of the number of items. Statistics of the class generation result is shown in Table 5.1.

5.4 Feature Extraction and Analysis

We approach the problem of predicting knowledge state ($k(t_j)$) and knowledge gain ($\Delta k(t_i, t_j)$) described in Section 5.2 with supervised models for classification, where details about the applied classification models are given in Section 5.5.1. To this end, each session s is represented by a feature vector $\vec{v} = (f_1, f_2, ..., f_n)$, where considered features are described in Section 5.4.1 and analyzed in Section 5.4.2.

5.4.1 Features Considered

We extracted features according to multiple dimensions of a search session, structured into five categories, namely features related to the *session*, *queries*, *SERP*, *browsing* behavior and *mouse* movements. The SERP category consists of features extracted from direct interactions with SERP items, while the browsing category consists of features extracted from subsequent user navigation beyond simple SERP clicks. The majority of features is motivated by existing literature, yet none of the features have been used on the inferential tasks of this work.

All considered features f_i are listed in Table 5.2 together with the Pearson Correlation Coefficient scores $Corr(f_i, \Delta k(t_i, t_j))$, $Corr(f_i, k(t_j))$ between the respective feature and the knowledge gain (state).

Table 5.2 Features for prediction of knowledge gain and knowledge state.

	Notation	$Corr(f_i, \Delta k(t_i, t_j))$	$Corr(f_i, k(t_j))$	Feature description
Session	$s\_duration$	-0.020	0.066	Duration of the search session of a worker on a given topic
	$s\_duration\_per\_q$	-0.019	0.066	Session duration per query
Query	q_num	0.052	0.103	Number of queries in session s
	$q\_term\_\{max, min, avg, total\}$	{0.0002,-0.094, -0.042,0.047}	{0.065,0.032, 0.051,0.068}	Maximum, minimum, average, total number of query terms
	$q\_uniq\_term\_\{max, min, avg, total\}$	{0.016,-0.087, -0.024,0.06}	{0.104,0.05, 0.084,0.089}	Maximum, minimum, average number of unique terms per query
	$q\_uniq\_term\_ratio$	0.083	-0.002	Number of query terms / unique query terms ($\frac{q\_uniq\_term\_total}{q\_term\_total}$)
	$q\_len\_\{first, last\}$	{-0.049,0.055}	{0.031,0.105}	First, last query length
	$q\_uniq\_term\_\{first, last\}$	{-0.023, -0.040}	{0.036,0.087}	Number of unique terms of first, last query
	$q\_complexity\_\{max, min, avg\}$	{0.097,0.086, 0.093}	{0.087,0.078, 0.049}	Maximum, minimum, average of query complexity
	$q\_complexity\_max\_diff$	{0.092}	{0.077}	Difference between the maximum and minimum complexity
SERP	$SERP\_click$	-0.009	0.063	Total number of click on search result
	$SERP\_click\_rank\_\{highest, lowest, avg\}$	{-0.101,-0.021, -0.017}	{-0.063,0.047, 0.095}	Average, highest, lowest rank of the clicks
	$SERP\_click\_interval$	0.036	0.022	Average interval between clicks
	$SERP\_click\_per\_query$	-0.007	-0.012	Average number of clicks per query
	$SERP\_no\_click\_query\_\{num, pct\}$	{0.041,-0.051}	{0.077,0.029}	Number, percentage of SERP with no clicks
	$SERP\_time\_\{total, avg, max\}$	{0.039,0.022, 0.049}	{0.091,-0.008, 0.043}	Total, average, maximum time spend on SERPs
	$SERP\_avg\_time\_to\_first\_click$	-0.002	-0.027	Time till first click
Browsing	b_num	-0.018	0.075	Total number of pages browsed in session
	b_uniq_num	0.029	0.109	Number of unique pages browsed in session
	$b\_num\_per\_q$	-0.017	-0.016	Average number of page browsed per query
	$b\_uniq\_num\_per\_q$	-0.017	-0.016	Average number of unique page viewed per query
	b_time_total	0.243	0.134	Total active time on the pages
	$b\_time\_avg\_per\_q$	0.236	0.063	Average active time on the browsed pages per query
	$b\_time\_\{max, avg\}\_per\_page$	{0.306,0.291}	{0.104,0.089}	Maximum, average active time on the browsed pages
	$b\_revisited\_ratio$	-0.058	-0.020	Ratio of revisited pages
	$b\_\{num, pct\}\_from\_SERP$	{-0.017,0.058}	{0.074,0.056}	Number, percentage of pages visited through SERP
	$b\_\{num, pct\}\_from\_non\_SERP$	{-0.056,0.057}	{-0.028,0.025}	Number, percentage of pages visited through pages other than SERP
	$b\_distinct\_domain\_num$	-0.033	0.102	Number of distinct domains of the visited pages
	$b\_ttl\_len\_\{max, min, avg, total\}$	{-0.08,-0.058, -0.078,-0.109}	{0.146,0.106, 0.146,0.082}	Maximum, minimum, average, total page title length
	$b\_page\_size\_\{max, min, avg, total\}$	{0.109,-0.093, 0.122,0.086}	{-0.055,-0.074, -0.057, -0.01}	Maximum, minimum, average, total page size
	$b\_ttl\_q\_overlap\_\{max, min, avg, total\}$	{0.15,0.089, 0.14,0.091}	{0.005,-0.028, -0.018,0.023}	Maximum, minimum, average and total overlap between query and page title
	$b\_url\_q\_overlap\_\{max, min, avg, total\}$	{0.16,0.064, 0.133,0.044}	{0.041,0.018, 0.025,0.028}	Maximum, minimum, average, total term overlap between query and page URL
Mouse	m_num	0.066	0.113	total number of mouseovers in the session
	$m\_num\_per\_q$	0.094	0.053	average number of mouseovers per query
	m_rank_max	0.091	0.067	max mouseover rank in the session
	$m\_rank\_max\_per\_q$	0.095	0.039	average max mouseover rank per query
	m_scroll_dist	0.120	0.058	total scroll distance in session
	$m\_scroll\_dist\_per\_q$	0.120	0.025	average scroll distance per query
	$m\_scroll\_max\_pos$	0.142	0.052	max scroll position in session
	$m\_scroll\_max\_pos\_per\_q$	0.127	0.021	average max scroll position per query

Session Features. The relation between feature *s_duration* and different stages of learning has been discussed by Jansen et al. [JBS09]. It has been shown that there is a difference in the duration of sessions among the classifications in Anderson and Krathwohl's taxonomy [AKA+01]. White et al. [WDT09] also found that the sessions conducted by domain experts were generally longer than non-expert sessions.

Query Features. Several prior works [JBS09, Arg14, WDT09] have investigated the correlation between query activities in a search session and learning performance. Based on the study by White et al. [WDT09], the *number of queries* (*q_num*) applied by experts and non-experts show big differences across domains: non-expert users usually run significantly more queries than experts. Jansen et al. [JBS09] also found that the *number of queries* applied on learning tasks classified as *applying stage* was significantly different from other l*earning stages*.

The *length of queries* ($q\_term\_max\{min, avg, total\}$) has been found to have a strong correlation with learning outcome by Zhang et al. [ZCB11]. Their study shows that the *average query length* and user domain knowledge is correlated with a Pearson correlation score of 0.344.

The *complexity of queries* ($q\_complexity\_max\_diff$) has been investigated by Eickhoff et al. [ETWD14], and has been found to evolve during the learning process. We applied the same query complexity measure as in [ETWD14], which is computed based on the dictionary created by Kuperman et al. [KSGB12] that contains a listing of more than 30,000 English words along with the age at which native speakers typically learn the term. The maximum age of acquisition across all query terms is used as query complexity.

Furthermore, the investigation from Arguello [Arg14] shows that beside the number of total terms, the *number of unique terms* ($q\_uniq\_term\_\{max, min, avg, total\}$, $q\_uniq\_term\_ratio$) in the session is strongly correlated with knowledge level on the task, while the number and ratio of stop words do not have a big difference when comparing between search sessions with different levels of domain knowledge.

As we aim at predicting knowledge state change during a session, similarly to the features discussed above, we extract the features $q\_len\_\{first, last\}$ and features $q\_uniq\_term\_\{first, last\}$, which potentially are indicators of the knowledge level at the beginning and end of the session.

SERP Activity Features. Some activities on SERP have also been investigated by previous works. Specifically, Collins-Thompson [CTRHS16] found that the *total number of clicks on SERP* ($SERP\_click$) is strongly correlated with a user's understanding of the topic. The analysis shows that users tend to click more often when having stronger interest in the topic.

The *ranking position* of the clicked URL on SERP has also been shown to be a strong indicator of user domain knowledge by Zhang et al. [ZCB11]. In [Arg14], the authors discovered that the difficult tasks with which a user is less knowledgeable are associated with more clicks ($SERP\_click$), more clicks on lower ranked doc-

uments ($SERP\_click\_rank\_\{highest, lowest, avg\}$), more abandoned queries without any click ($SERP\_no\_click\_query\_\{num, pct\}$), i.e. queries without clicks, longer *time till first click* ($SERP\_avg\_time\_to\_first\_click$) and longer *time till next click* ($SERP\_click\_interval$).

Browsing Features. Browsing features such as *number of documents viewed* (b_num, b_uniq_num) and *average number of documents viewed per query* ($b\_num\_per\_q$, $b\_uniq\_num\_per\_q$) were shown by several previous works [ETWD14, JBS09, Arg14, GS06] to be positively correlated with the knowledge improvement. More detailed features corresponding to the browsing behavior have also been studied, indicating that the more difficult a task is for a user, the higher the *ratio of revisited pages* ($b\_revisited\_ratio$) is.

Despite the number of pages visited, the time spent (corresponds to features b_time_total, $b\_time\_\{max, avg\}\_per\_q$ etc.) on the accessed pages are found to vary to a large extent between domain expert and non-expert [WDT09]. Feature $SERP\_time\_\{total, avg, max\}$ was shown to be effective for predicting the user's assessment of task difficulty [Arg14], which is subject to the user's knowledge state.

We further distinguish the viewed pages into two sets {pages navigated through SERP, pages navigated through non-SERP}, by parsing its ancestor page. Hence we extract the features $b\_\{num, pct\}\_from\_SERP$ and $b\_\{num, pct\}\_from\_non\_SERP$ that are motivated by the features introduced above.

The content of the accessed Web documents strongly influence the user's learning outcome. White et al. [WDT09] found that domain experts encountered different and more diverse domains (feature $b\_distint\_domain\_num$) than domain novices. Several other document content related features: *page size* ($b\_page\_size\_\{max, min, avg, total\}$), *title length* $b\_ttl\_len\_\{max, min, avg, total\}$ have also been found to evolve during the learning process [ETWD14]. Based on the assumption that domain experts and novices have different capabilities of choosing learning resources, for instance, experts are able to recognize useful documents without query terms presented in the page title, we computed features based on the overlap between page title and query ($b\_ttl\_q\_overlap\_\{max, min, avg, total\}$). The page URL as a complementary source containing hints about a page's content has also been considered in the feature extraction process ($b\_url\_q\_overlap\_\{max, min, avg, total\}$).

Mouse Features. Features in the *Mouse* category are indicators of quantity and quality of user interactions with a knowledge source and were also shown to be effective for predicting the user's assessment of task difficulty [Arg14].

5.4.2 Feature Analysis and Selection

As a basis for feature selection, we analyze the features with respect to their relationship to knowledge gain and knowledge state, as well as their redundancy.

Correlation between feature and KG (KS). In order to select the most influ-

ential features for the prediction task, we compute the Pearson correlation coefficient between each feature and the knowledge gain (knowledge state), i.e. $Corr(f_i, \Delta k(t_i, t_j))$ and $Corr(f_i, k(t_j))$. The correlation scores are shown in Table 5.2. Based on the computed score, we select the features fulfilling the condition $Corr(f_i, \Delta k(t_i, t_j)) \geq \beta$ for the knowledge gain prediction task and $Corr(f_i, k(t_j)) \geq \gamma$ for the knowledge state prediction task. Performance of the prediction model using features selected based on varied β and γ has been evaluated and corresponding results are presented in Section 5.6.

Correlation between features. We compute the Pearson correlation coefficient between each pair of features $Corr(f_i, f_j)$. If $Corr(f_i, f_j) \geq \tau$, i.e. features appear to be not independent, we remove the feature from the feature set, that has lower $Corr(f_i, \Delta k(t_i, t_j))$ respectively lower $Corr(f_i, k(t_j))$ for knowledge gain (state) prediction. We evaluate the performance of the prediction model for different values of τ. The feature selection results are reported in Section 5.6.

5.5 Evaluation - Experimental Setup

This section gives the details of the tested configurations, the baseline approach that we compared against in the experimental evaluation and the metrics used in the evaluation.

5.5.1 Configurations and Parameters

We experimented with a large number of configurations for each of our prediction models, as listed below:

- **Classifier.** We apply a range of standard models for the classification of the knowledge gain and knowledge state, namely, Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Random Forrest (RF), and Multilayer Perceptron (MP). For our experiments, we used the Weka library for Java[2]. For each of the configurations described below, we perform grid search to tune the hyperparameters of all of the classifiers. In Section 5.6, we report the result of the best performing hyperparameter configuration for each classifier.

- **β (γ)- threshold for feature selection based on correlation between feature and KG (KS).** We compare prediction performance before and after applying the selection based on feature-KG (KS) correlation. We set the threshold β (γ) for selecting the features in the range of {0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3} ({0.0, 0.05, 0.1, 0.15}). We omit results for larger β (γ), as the resulting number of features is insufficient for training a classifier.

[2] https://www.cs.waikato.ac.nz/ml/weka/

- τ **- threshold for feature selection based on correlation between features.** We also experimented with different τ in the range of {1.0, 0.95, 0.9, 0.85, 0.8, 0.75, 0.7}. The number of features in the feature set corresponding to given τ, β and γ is reported in Table 5.3.

Table 5.3 Number of features of different configurations.

	β (**KG**)							γ (**KS**)			
	0.0	0.05	0.1	0.15	0.2	0.25	0.3	0.0	0.05	0.1	0.15
$\tau = 1.0$	70	43	16	6	4	2	1	70	41	11	0
$\tau = 0.95$	66	42	16	6	4	2	1	66	38	11	0
$\tau = 0.9$	56	37	13	6	4	2	1	58	34	11	0
$\tau = 0.85$	43	29	10	6	4	2	1	44	24	9	0
$\tau = 0.8$	39	26	7	4	2	1	1	38	19	8	0
$\tau = 0.75$	37	25	7	4	2	1	1	34	17	7	0
$\tau = 0.7$	33	24	6	3	1	1	1	32	16	7	0

5.5.2 Baseline

As discussed in Section 5.1, the tasks addressed in this work are comparably novel. To the best of our knowledge, there are no existing baselines for the task of knowledge gain prediction during informational Web search missions. Therefore, we compare our approach for a number of configurations (described above), using multiple standard classification models. For the prediction of knowledge state $k(t_j)$, we compare our approach in addition to one existing baseline [ZCB11]. KS_{Zhang} refers to the linear regression model fitted by Zhang et al. [ZCB11] for domain knowledge prediction as shown in Equation 5.1.

$$KS_{Zhang} = -1.466 + 0.039 \cdot Saved + 0.147 \cdot Q_{len} + 0.130 \cdot Rel_{mean} \tag{5.1}$$

Saved represents the number of documents saved by the user, which is an extremely sparse feature in a real search environment and does not appear in our dataset. Q_{len} is the mean query length and Rel_{mean} is the mean rank of documents opened in SERPs. As the output of the baseline regression model is a real number, we convert the result into 3 classes according to the definition given in Section 5.3.

5.5.3 Evaluation Metrics

For both tasks, we run repeated 10-fold cross-validation with 10 repetitions on all the approaches and configurations described in Section 5.5.1 and evaluate the results according to the following metrics:

- **Accuracy (*Accu*) across all classes:** percentage of search sessions that were classified with the correct class label.
- **Precision (P), Recall (R), F1 ($F1$) score of class i:** we compute the standard precision, recall and F1 score on the prediction result of each class i.
- **Macro average of precision (P), recall (R), and F1 ($F1$):** the average of the corresponding score across 3 classes.
- **Runtime:** the time consumed for completing the 10-fold cross-validation on experimental dataset in milliseconds.

To analyze the usefulness of individual features, we make use of the *Mean Decrease Accuracy (MDA)* metric, which is based on the Random Forest model, i.e. a very well performing model for both tasks as shown in Section 5.6. MDA quantifies the importance of a feature by measuring the change in prediction accuracy of the Random Forest model when the values of the feature are randomly permuted compared to the original observations.

5.6 Results: Prediction Performance and Feature Analysis

In this section, we report the evaluation results of the prediction performance as well as an analysis of feature importance.

5.6.1 Knowledge Gain Prediction

Table 5.4 Performance in knowledge gain prediction task.

Method	τ	β	#Feature	Runtime	Low			Moderate			High			Macro average			All
					P	R	F1	P	R	F1	P	R	F1	P	R	F1	Accu
NB	≥0.85	0.25	2	19.1	0.450	0.747	0.562	0.483	0.268	0.344	0.513	0.384	0.439	0.482	0.467	0.448	0.469
LR	0.85	0.05	29	653.9	0.498	0.537	0.516	0.459	0.382	0.416	0.379	0.431	0.403	0.445	0.450	0.445	0.450
SVM	0.90	0.00	56	441.6	0.488	0.595	0.536	0.487	0.340	0.400	0.410	0.469	0.437	0.462	0.468	0.458	0.465
RF	0.95	0.00	66	3739.3	0.521	0.542	0.531	0.469	0.410	0.437	0.425	0.480	0.450	0.472	0.477	**0.473**	**0.475**
MP	≥0.85	0.25	2	1919.3	0.452	0.556	0.497	0.421	0.312	0.356	0.425	0.450	0.435	0.433	0.439	0.429	0.435

Performance of different Configurations. For each of the 245 distinct configurations described in Section 5.5, we run repeated cross-validation as described in the previous section.

From all the different combinations of τ and β as listed in Table 5.3, we present the result of the configuration that produces the highest accuracy for each classifier in Table 5.4. A complete set of the evaluation results are available online[3]. We observed

[3] https://sites.google.com/view/predicting-user-knowledge

that in the knowledge gain prediction task, the highest average F1 score across classes and the highest accuracy always appear in the same configuration for all the classifiers except for LR, where there is a minor difference of .001 in F1 between the two.

The *Random Forest* classifier achieves the best performance in terms of accuracy and average F1 score, slightly outperforming Naive Bayes and SVM. As shown, Naive Bayes is the most efficient classifier in terms of computation time for feature sets of comparable size. Comparing classification performance for the different classes, the results for each of the classifiers consistently show higher F1 scores for the 'Low' and 'High' knowledge gain classes than for the 'Moderate' knowledge gain class. Overall, prediction accuracy is above 0.435 and the F1 score is above 0.429 for all of the classifiers, which indicates that the set of features we extracted from search activities can provide meaningful evidence for predicting knowledge gain.

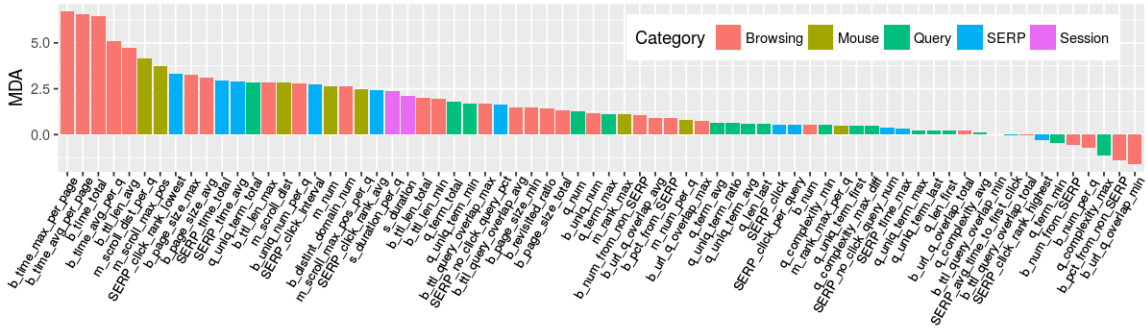


Figure 5.1 Feature importance for knowledge gain prediction.

Feature Impact. The MDA results of each feature are shown in Figure 5.1. Based on the result, the most important features are: *b_time_max_per_page*, *b_time_total* and *b_time_avg_per_page*. Mostly active time related features which reflect the effort users spend on learning. Similarly, these features are immediately followed by *m_scroll_dist_per_q*, *m_scroll_max_pos*, and *SERP_click_rank_lowest*, three features that are indicators for the amount of information a user has seen during the session.

As shown in Table 5.4, the NB and MP classifier performs best when using 2 features only, namely *b_time_max_per_page* and *b_time_avg_per_page*. These two features are confirmed as the most important features by our feature importance analysis.

Regarding feature categories, the 10 most useful features in terms of MDA belong to the browsing, mouse and SERP categories. Surprisingly, although multiple query features had above average correlation to knowledge gain (see Table 5.2) – in particular, the features related to query complexity ($q\_complexity\_\{max, min, avg\}$ and $q\_complexity\_max\_diff$) had correlations ranging from .086 to .097, compared to the median of .042 – *q_uniq_term_total* is the only query feature among the 25 highest ranked according to MDA. Analogously, both session features *s_duration* and *s_duration_per_q* appear among the 25 highest ranked features despite their relatively low correlations of -.02 and -.019.

5.6.2 Knowledge State Prediction

Table 5.5 Performance in knowledge state prediction task.

Method	τ	γ	#Feature	Runtime	Low			Moderate			High			Macro average			All
					P	R	F1	P	R	F1	P	R	F1	P	R	F1	Accu
NB	≤0.75	0.1	7	23.5	0.352	0.712	0.470	0.424	0.218	0.287	0.370	0.211	0.268	0.382	0.380	0.342	0.369
LR	1.00	0.05	41	797.7	0.338	0.383	0.359	0.402	0.368	0.384	0.372	0.359	0.366	0.370	0.370	0.370	0.370
SVM	0.95	0.05	38	292.3	0.359	0.479	0.409	0.395	0.303	0.342	0.409	0.386	0.397	0.388	0.389	0.383	0.385
RF	1.00	0.00	70	4023.4	0.443	0.456	0.449	0.394	0.358	0.374	0.418	0.447	0.432	0.418	0.421	**0.418**	**0.418**
MP	1.00	0.05	41	43619	0.380	0.414	0.396	0.398	0.298	0.341	0.385	0.461	0.419	0.388	0.391	0.385	0.387
KS_{Zhang}	-	-	2	23	0.320	0.428	0.366	0.328	0.240	0.277	0.362	0.355	0.359	0.337	0.341	0.334	0.335

Performance of different Configurations. We have experimented with all different combinations of τ and γ as listed in Table 5.3 for all considered classifiers. The result of the configuration that produces the highest accuracy for each classifier is shown in Table 5.5. We observe that in the knowledge state prediction task, the highest average F1 score across classes and the highest accuracy always appear in the same configuration for all the classifiers except Naive Bayes (average F1 of the highest accuracy configuration is 0.006 lower than the maximum average F1).

Among all evaluated classifiers, Random Forest reaches the highest accuracy and F1 score, outperforming the other classifiers.

Comparison to Baseline. We compare the performance of our approach against the baseline method (KS_{Zhang}), shown in the last row in Table 5.5. The result suggests that, the linear regression model fitted in previous work based on data collected through a lab study does not perform well in the knowledge state prediction task and is outperformed by all five classifiers following our approach.

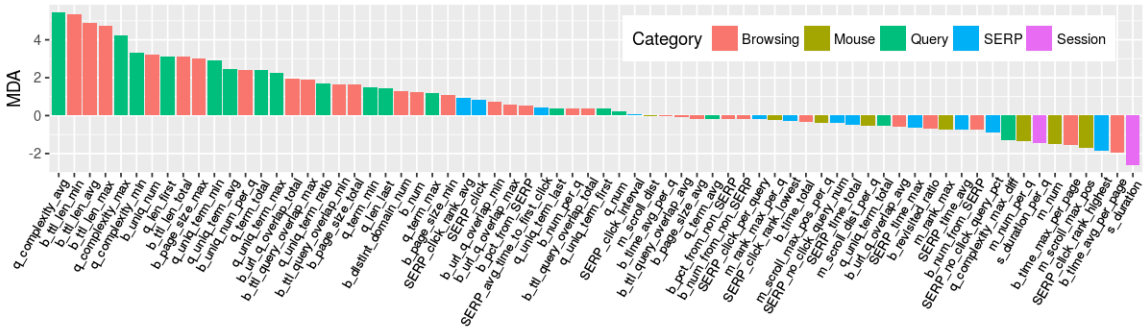


Figure 5.2 Feature importance for knowledge state prediction.

Feature Impact. The MDA results of each feature in the knowledge state prediction tasks are shown in Figure 5.2. The most important features (*q_complexity_avg*, *b_ttl_len_min* and *b_ttl_len_avg*) reflect the user's capability of constructing a query and choosing relevant resources. In terms of feature categories, all of the highest ranked features for this task belong to the query and browsing categories.

Compared to the knowledge gain task, query complexity features are considerably more useful ($q\_complexity\_\{min, max, avg\}$), while features related to time and effort invested, like $b\_time\_max\_per\_page$ and $b\_time\_avg\_per\_page$, are among the lowest ranked. Other query features related to the used vocabulary (e.g. $q\_uniq\_term\_min$, $q\_uniq\_term\_avg$, and q_term_total) are ranked similarly highly. Apparently, while the time taken by users to take in the discovered documents is predictive of their knowledge gain, their capability of using complex queries and selecting relevant resources reveals more about their knowledge state.

5.7 Discussion

Based on the experimental results, we conclude that: i) knowledge gain (state) can be predicted during informational search sessions with a certain level of accuracy, ii) performance of the knowledge gain prediction appears to be generally better, suggesting that the task is easier given the nature of our data, and iii) the performance of the prediction approach is better for more extreme classes, i.e. for low and high knowledge gain (state) classes, whereas performance on the moderate classes is lowest in both tasks, presumably due to the moderate classes being the most overlapping ones with respect to their characteristics. In this section we discuss some of the reasons behind these observations.

Most of the features we considered were found to correlate rather weakly with knowledge gain (state). Intuitively, this could be due to the limited duration of the search sessions (just over 5 minutes on average). This could potentially reduce the predictive power of certain features, such as the number of queries or the number of accessed documents. This also rendered evolution-oriented features, which would capture the evolution of queries and behavior throughout a session predictively poor. While these would supposedly be highly indicative of the knowledge gain, they require longer sessions than are usually observable in real-world search sessions as well as in our experimental data.

For the prediction of knowledge gain, our feature analysis result shows that the most important features are the ones related to the user's active time. As our experimental dataset contains mostly short sessions, it is understandable that the time spent affects the knowledge gain strongly. However, we believe that in longer search sessions, the learning pattern and the initial knowledge state of a user might be more influential for the knowledge gain than in short sessions. Further experiments are required to establish this.

The results suggest that with the presented approach, the knowledge gain prediction is an easier task than the knowledge state prediction. As shown in Figure 5.2, the most important features for knowledge state prediction are the features related to the content of queries and browsed documents. Intuitively, these features are also central to the knowledge gain prediction task. Yet, we observe that although the topic

descriptions that were given to the users typically provided central keywords for the first query, only a very limited set of queries (1-2) are fired by most users. Given the small number of queries in each session, the query features are less distinguishable and hence, less indicative of the knowledge gain. Thus, query evolution is observable only to a very limited extent.

6 Topic-independent Modeling of User Knowledge in Informational Search Sessions

The work in this chapter continue to focus on the topic of improving the knowledge assessibility on the Web for human. In Chapter 4 and 5, we investigated the relation between user interaction and their knowledge on the search task, and presented an approach for the prediction of knowledge state as well as knowledge gain of a user using a range of behavioral signals captured during online search sessions. The findings demonstrate that knowledge gain/state can be predicted from user behavior throughout search sessions. However, the initial attempt in Chapter 5 has been constrained by limited feature sets. Insights into the generalizability of predictive models across topics are still shallow.

Building on such prior works, this work introduces a novel set of Web resource-centric features and investigates their impact on the knowledge gain/state prediction task. Web resource features consider characteristics of resources a user interacts with, such as their linguistic tone, their complexity or structural aspects of an HTML page. We make valuable contributions given that reliable training data for such tasks is sparse and costly to obtain. The feature space of potentially relevant features is large: 179 distinct features (109 web resource features, 70 user behavior features) are investigated in total in our work. Thus, we introduce various feature selection strategies geared towards selecting a limited subset of effective and generalizable features by considering feature correlation with knowledge gain/state, topic-dependency of feature performance and feature redundancy.

The supervised models that we propose in this work outperform the state-of-the-art and show an average F1-score improvement by 25.5%, and an increase in accuracy by 23.2% on average across different prediction tasks. In summary, our contributions include the following:

- **Novel feature sets.** We introduce and experimentally evaluate novel Web resource feature sets (109 features in total) for the task of knowledge state (KS) and knowledge gain (KG) prediction, which extend state of the art models.

- **Feature analysis.** We conduct comprehensive feature analysis assessing both generalisability of features across search topics as well as their overall effectiveness in the aforementioned prediction tasks. Findings from this analysis can inform future work for user modeling in search sessions in various ways. Moreover, our analysis can be leveraged to build computationally efficient models through a limited set of effective features.
- **Feature selection approach.** In order to cope with the wide variety and large number of features in the presence of very sparse training data, we introduce a novel approach for feature selection which combines feature correlation with target variables (KG/KS) as well as the topic-dependency of feature performance. By doing so, we identify best performing features in cross-topic prediction settings and facilitate generalisable models.
- **Improved prediction models.** We evaluate our features and feature selection approach by building supervised classifiers which outperform state-of-the-art baselines for the knowledge gain/state prediction on *unseen* topics. On average, our improved models outperform the previous state-of-the-art baseline [YGH+18] by 20.6%, 39.9%, and 16% (average F1 score) in the tasks of knowledge gain, pre-knowledge state, and post-knowledge state prediction, respectively.

Potential applications of our work include the consideration of a user's knowledge state during the retrieval and ranking step as part of state-of-the-art Web search. Our findings are equally relevant for the classification and guidance of search behavior in learning-oriented search scenarios, for example, in class rooms, libraries or work environment.

6.1 Tasks & Dataset

6.1.1 Tasks

We reuse the definition of *intentional learning-related search session*as defined in Definition 2. We refer to such intentional learning-related *search session* as "session" in the remainder of this chapter for simplicity.

Let s be a search session starting at time t_i and ending at time t_j aimed at satisfying a particular information need, that is, a learning intent ι of user u. In this work, we study the knowledge indicators (KIs): pre-knowledge state (pre-KS) $k(t_i)$, post-knowledge state (post-KS) $k(t_j)$ and knowledge gain (KG) $\Delta k(t_i, t_j)$ during time period $[t_i, t_j]$. This work aims at extending the understanding of user knowledge (change) in the informational search process and build topic independent models (with respect to users' learning intents), to predict the aforementioned knowledge indicators. More specifically, this work addresses the following tasks:

[T1] Understanding the relation between Web resource features and a user's knowledge state and knowledge gain. The features we considered are described in

Section 6.2.

[T2] Understanding the topic-specificity of individual features, i.e. dependency between feature performance and information needs (topics), investigating feature selection strategies geared towards selecting effective and topic-independent features for modeling KIs. The investigated features include the document features described in Section 6.2, as well as the user interaction features studied by previous works [GYDH18, YGH+18].

[T3] Build generalizable models that can be used in real-world search environments on unseen topics for predicting the KIs. We aim to classify a specific KI, e.g. knowledge gain $\Delta k(t_i, t_j)$, with respect to a particular information need. For the sake of this work, a user's knowledge state is defined by the user's capability to correctly respond to a set of questions about the corresponding information need. A user's knowledge gain is defined as the improvement of user capability (accuracy) to correctly respond to a set of test questions about the corresponding information need. The classification goal is introduced in Section 6.1.2.

6.1.2 Dataset

For the analysis and experimental evaluation, we use the dataset as described in Section 4.2. Furthermore, we rerun the study following the same procedure to collect more search sessions, this finally result in 1100 search sessions correspond to 11 different topics (100 sessions for each topic). To ensure reliability of responses and the resulting behavioral data logged during the search sessions, we filtered sessions using criteria introduced in Section 4.2.3 and an additional heuristic, which is filtering out workers who interacted with non-English resources. The rationale behind considering this filter was that many of the features (see Section 6.2) we extracted from the Web resource content rely on certain dictionaries, which are currently only available for the English language.

After applying all the aforementioned filters, we retain 233 search sessions, with 1.361 queries and 2.622 clicks per session on average. Figure 6.1 shows the number of Web search sessions corresponding to each topic. The topic "HIV" has 31 sessions, i.e. the largest number of sessions. The topic "NASA" has only 15 sessions. The mean number of Web search sessions for each topic is 21.18.

Knowledge State and Knowledge Gain Classes. For the classification tasks described in [T3], we follow the same approach as used in previous work [YGH+18], i.e. a *Standard Deviation Classification* approach to obtain three classes of learners with regard to their level of knowledge. Assuming approximately normal distributions of the respective test scores (X) for the different topics, we transformed the test scores into Z-scores with a mean of 0 and a Standard Deviation (SD) of 1 (standardization). We then used statistically defined intervals (low: X $<$ 0.5 SD; moderate: -0.5 SD $<$ X

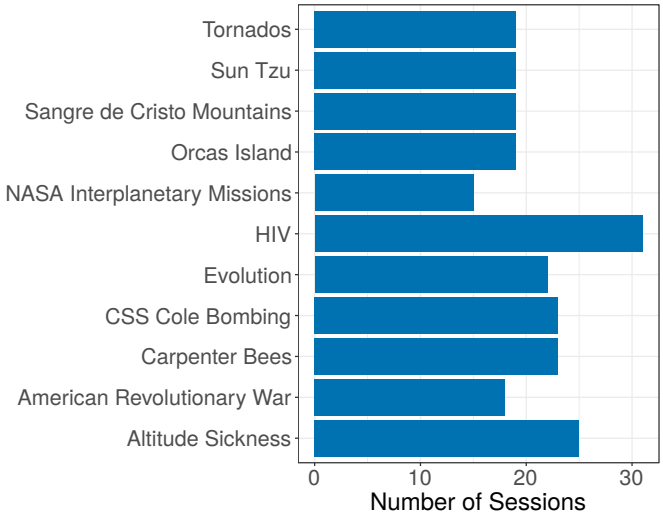


Figure 6.1 Number of Web search sessions pertaining to each topic and the associated information need after filtering.

$<$ 0.5 SD; high: 0.5 SD $<$ X) for the classification of the learners into roughly equal groups with low, moderate, or high pre-KS. The same procedure was repeated for post-KS and KG. Table 6.1 shows the resulting numbers of learners for the respective classes and underlying statistics.

Table 6.1 Knowledge state and knowledge gain classes created based on thresholds of *mean* $\pm$ *0.5SD*.

Task	Mean	SD	Low	Moderate	High
pre-knowledge state	0.36	0.255	87	52	94
post-knowledge state	0.66	0.174	61	95	77
knowledge gain	0.23	0.208	84	84	65

6.2 Feature Extraction

We approach the problem of predicting KI as described in Section 6.1.1 with supervised classification models, where details about the applied models are given in Section 6.3.1. Each session s is represented by a feature vector $\vec{v} = (f_1, f_2, ..., f_n)$; where the features considered are described in the following subsections.

6.2.1 Web Resource Features

We introduce 109 Web resource features in total. To ensure readability of the thesis, we list only a subset of features in this section. The full set of features are listed in

Table 6.2 Considered Web resource features and user behavior features (not complete), highlighted cells having $p\text{-}value \geq 0.05$.

		Corr			SDoC			
	Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
Complexity	c_adj_avg	-0.287	-0.329	0.076	0.166	0.230	0.187	Ratio of the number of adjectives to the total number of words
	c_aoa_avg	0.265	0.199	-0.157	0.227	0.177	0.245	Average age-of-acquisition rating of words in each webpage
	c_char_avg	-0.077	-0.084	0.024	0.180	0.187	0.193	Average number of characters per term
	c_fk_avg	-0.174	-0.203	0.043	0.236	0.155	0.284	Flesch-Kincaid Grade Readability Index
	c_gi_avg	-0.062	0.092	0.153	0.257	0.272	0.235	Gunning Fog Grade Readability Index
	c_noun_avg	-0.165	0.001	0.203	0.153	0.179	0.118	Ratio of the number of nouns to the total number of words
	c_oth_avg	0.117	-0.007	-0.149	0.143	0.150	0.105	Ratio of the number of *other words* to the total number of words
	c_sentence_avg	-0.024	-0.006	0.024	0.230	0.181	0.257	Average number of words per sentence
	c_smog_avg	-0.041	0.088	0.124	0.230	0.248	0.202	SMOG Readability Index
	c_uniq_word_avg	0.015	0.164	0.118	0.202	0.178	0.229	Ratio of the number of unique words to the total number of words
	c_verb_avg	0.342	0.234	-0.222	0.188	0.235	0.199	Ratio of the number of verb to the total number of words
	c_word_avg	-0.191	-0.243	0.030	0.196	0.199	0.190	Number of words in each webpage
HTML Structure	h_img_avg	-0.187	-0.273	0.001	0.204	0.191	0.191	Number of <img>elements
	h_link_avg	-0.099	-0.240	-0.079	0.190	0.221	0.192	Number of outbound links
	h_nav_ul_avg	0.206	0.184	-0.098	0.249	0.172	0.238	Number of elements embedded in <nav>elements
	h_oth_ul_avg	-0.243	-0.305	0.043	0.218	0.148	0.187	Number of elements not in <nav>elements
	h_p_avg	-0.281	-0.230	0.151	0.205	0.255	0.238	Average length of each paragraph in <p>elements
	h_script_avg	0.165	0.067	-0.145	0.221	0.141	0.245	Number of <script>elements
Linguistic	l_Analytic_avg	-0.469	-0.294	0.329	0.192	0.224	0.172	Number of analytic words
	l_anger_avg	-0.250	-0.151	0.179	0.172	0.210	0.090	Number of anger words
	l_bio_avg	0.564	0.364	-0.386	0.238	0.207	0.281	Number of biological process words
	l_body_avg	0.469	0.336	-0.293	0.236	0.186	0.264	Number of body words
	l_focuspresent_avg	0.514	0.302	-0.376	0.263	0.226	0.240	Number of present focus words
	l_health_avg	0.556	0.351	-0.387	0.186	0.217	0.259	Number of health words
	l_money_avg	-0.156	-0.026	0.169	0.130	0.156	0.188	Number of money words
	l_relativ_avg	-0.264	-0.295	0.077	0.226	0.285	0.209	Number of relativity words
	ls_article_avg	-0.297	-0.294	0.118	0.187	0.170	0.174	Number of articles
	l_percept_avg	-0.043	-0.039	0.020	0.240	0.172	0.316	Number of perceptual processes
	ls_conj_avg	0.422	0.184	-0.362	0.208	0.182	0.182	Number of conjunctions
	ls_Dic_avg	0.398	0.164	-0.349	0.213	0.197	0.243	Number of dictionary words
	ls_number_avg	-0.358	-0.227	0.248	0.240	0.238	0.246	Number of numbers
	ls_Quote_avg	-0.339	-0.320	0.147	0.208	0.158	0.260	Number of quotation marks
	ls_you_avg	0.473	0.337	-0.297	0.155	0.247	0.201	Number of you pronouns
User Behavior	b_revisited_ratio	0.002	-0.043	-0.038	0.292	0.180	0.357	Ratio of revisited pages
	b_time_avg_per_q	-0.144	0.035	0.205	0.306	0.257	0.264	Average active time on the browsed pages per query
	b_ttl_len_avg	0.397	0.281	-0.251	0.203	0.190	0.258	Average page title length
	m_rank_max	-0.027	0.110	0.126	0.194	0.187	0.315	max mouseover rank in the session
	m_scroll_dist	-0.030	-0.011	0.028	0.278	0.220	0.353	total scroll distance in session
	q_len_first	0.033	0.002	-0.039	0.180	0.298	0.197	First query length
	q_uniq_term_first	0.023	0.019	-0.012	0.179	0.283	0.184	Number of unique terms of first query
	q_uniq_term_ratio	-0.185	-0.085	0.154	0.292	0.304	0.234	$\frac{\text{Number of unique query terms}}{\text{number of query terms}}$
	s_duration	0.113	0.087	-0.066	0.380	0.262	0.351	Duration of the search session of a worker on a given topic
	s_duration_per_q	0.113	0.087	-0.065	0.312	0.234	0.288	Session duration per query

Appendix A.

Document Complexity Features. The assumption behind the document complexity related features is that, the higher a user's knowledge state is on a topic, it is more likely that the user prefers documents with higher complexity. As previously reported [ETWD14], the number of words (*c_word*) can be an indicator for content complexity. Moreover, long words (*c_char*) are more likely to be specific and indicative of complex vocabularies than short words. Similarly, long sentences (*c_sentence*) have been found to indicate higher resource complexity than short sentences.

The syntactic structure of a document, which is represented by the ratio of the number of nouns, verbs, adjectives, or *other words* (i.e. words that are not verb, noun or adjective) to the total words ($c\_\{noun, verb, adj, oth\}$), is likely to suggest the intention and complexity of its content [HCTCE07].

The grade level readability index can be used to measure the readability of a document by computing a score based on the number of the syllables in words. The assumption is that it requires a higher education level to read a document with a higher score [HA17]. We compute three different readability grades: Gunning Fog Grade[1] (c_gi), SMOG [LH69] (c_smog) and Flesch-Kincaid Grade [KFJRC75] (c_fk). Using the age-of-acquisition (AoA) dictionary proposed by Kuperman et al. [KSGB12] that contains a listing of more than 30,000 English words along with the age at which native speakers typically learn the term , we compute the age-of-acquisition across all words on webpages (c_aoa), which provides another indicator of document complexity.

HTML Structural Features. Previous works [SCT18] have investigated the influence of images on user's learning outcome in Web search, they found a positive correlation between a relevant image and KG, and a negative correlation between the total number of images and KG. Here we do not distinguish between the relevant and irrelevant images due to the current lack of an automated approach that can be applied in a real-world search environment. We hence compute the number of <*img*> elements on webpages to estimate the number of images (h_img) it contains.

Prior work [DL07] found a negative association between the number of hyperlinks embedded in a webpage and the user's KG. The assumption is that people may not focus on the content in the presence of too many embedded links. We quantify the number of outbound links by counting the <*a*> elements (h_link). The average length of each paragraph (h_p) is one of the indicators of the required effort for understanding the resource [DL07]. The <*ul*> elements embedded (h_oth_ul) are often used to present important ideas of the document as an unordered list in a more structured and easily digestible fashion, and thus, may have a positive impact on KG. The <*script*> element is used to define a client-side script (e.g. JavaScript). Based on our observation, different types of websites adopt different styles of using scripts, e.g. Wikipedia uses far fewer scripts than typical commercial websites. We assume that the presence of scripts might be correlated with the possibility of whether a website suits learning-oriented needs, and therefore correlates with KG. We compute the number of scripts (h_script) on a webpage to serve as a feature.

Linguistic Features. We make use of the 2015 Linguistic Inquiry and Word Count (LIWC) dictionaries[2] to compute the features in this category. According to previous work [HA17], the amount of words on webpages that are correlated with different psychological processes and basic sentiment can influence a learner's cognitive state. Based on this assumption, we extracted 56 features. To ensure the readability of the thesis, we only show the features in this list of features that are discussed in this chapter in Table 6.2, where notations begin with $l\_$ in the *Linguistic* category.

The stylistic features capture grammatical characteristics, text style, and syntax of a document [HA17]. The writing style could affect the readability of a learning resource and the engagement of readers. We compute 35 relevant features using the

[1] http://gunning-fog-index.com/
[2] http://liwc.wpengine.com/

LIWC dictionary. We show the features that are discussed in this chapter from this feature list (features in Table 6.2 with notations beginning with $ls\_$ in the *Linguistic* category). All features in this category are named (in Table 6.2) according to the label generated by LIWC dictionary.

6.2.2 User Behavior Features

Apart from the resource content-related features introduced above, we also consider the 70 user behavior-related features that were introduced in Chapter 5. The user interaction-related features were extracted according to multiple dimensions of a search session, namely features related to the session, queries, SERP, browsing behavior and mouse movements. The SERP category consists of features extracted from direct interactions with SERP items, while the browsing category consists of features extracted from subsequent user navigation beyond simple SERP clicks. We listed the user behavior features that are discussed in the remainder of this chapter in Table 6.2. We made the full list of user interaction features available in Appendix A.

6.2.3 Feature Selection Strategies

For the classification tasks, we consider all 109 resource content-related features and 70 user behavior-related features as described above, denoted as F. However, due to the difficulty of obtaining user knowledge assessment data, the scale of training/testing data is limited. Hence, feature selection is important for building reliable models, and in particular, to avoid overfitting. The goal of this step is to select a set of features $F' \subseteq F$ that can produce the best performing model for the prediction of a specific knowledge indicator. In this section, we introduce 3 feature selection strategies. For the sake of simplicity, we refer to all knowledge indicators, i.e. pre-KS $k(t_i)$, post-KS $k(t_j)$ or KG $\Delta k(t_i, t_j)$ as KI in the following.

$Corr(f_i, KI)$ – **Feature Effectiveness.** We compute the Pearson correlation coefficient between each feature and the knowledge indicator $Corr(f_i, KI)$ across all sessions. The correlation scores are shown in Table 6.2. To ensure effectiveness of features, we select features fulfilling the condition $Corr(f_i, KI) \geq \alpha$ for the prediction task. Performance of the prediction model using features selected based on varied α has been evaluated and corresponding results are presented in Section 6.4.

$SDoC(f_i, KI)$ – **Generalizability.** In order to measure the topic dependency of features, we compute the correlation between a feature and a knowledge indicator for each topic and measure the standard deviation (SD) of the correlation score across topics. The intuition is that a small standard deviation of the correlation between a feature and the respective KI is indicative for a topic independent relationship that may generalize to other topics as well. For simplicity, we will refer to this metric as $SDoC$ (Standard Deviation of Correlation) in the remaining of this chapter. The

computation of $SDoC$ of feature f_i is shown in Equation 6.1.

$$SDoC(f_i, KI) = \sqrt{\frac{\sum_{j=1}^{N} \left(Corr^{(j)}(f_i, KI) - \frac{\sum_{j=1}^{N} Corr^{(j)}(f_i, KI)}{N} \right)^2}{N-1}} \tag{6.1}$$

where N is the number of topics in the sample data set (here $N = 11$), $Corr^{(j)}(f_i, KI)$ is the correlation between f_i and KI on the sample data corresponding to topic j. To improve the generalizability of the model, we remove topic dependent features using the $SDoC$ metric, that is, we keep only the features with $SDoC(f_i) < \beta$ on the respective knowledge indicator.

$Corr(f_i, f_j)$ – **Feature Redundancy.** We compute the Pearson correlation coefficient $Corr(f_i, f_j)$ between each pair of features in the feature set. If $Corr(f_i, f_j) \geq \tau$, i.e. features do not appear to be independent from each other, we remove the feature of the feature pair which has a lower $Corr(f_i, KI)$ for the corresponding prediction.

6.3 Experimental Setup

6.3.1 Approach Configurations & Baseline

Classifier. We apply a range of standard models for the classification tasks, namely, Naive Bayes (nb), Logistic Regression (lr), Support Vector Machine (svm) and Random Forest (rf). For our experiments, we used the *scikit-learn* library for Python[3]. We tune hyperparameters of the algorithms using grid search within the repeated cross-topic validation setup described in Section 6.3.2.

Feature Category. In order to evaluate the influence of resource content-related features on the prediction of KIs, we compare between the performance of the prediction models using: 1) only user behavior features (feature category UB), 2) only Web resource features (feature category WR) and 3) using both user behavior and Web resource features (feature category $WR\&UB$).

Feature Selection Strategy. We test a range of thresholds for selecting the features according to the strategies introduced in Section 6.2.3. Specifically, for the feature selection based on $Corr(f_i, KI)$, we apply $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$; for the selection based on $SDoC(f_i, KI)$, we apply $\beta \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$; for the selection based on $Corr(f_i, f_j)$, we apply $\tau \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. The thresholds α, β, τ are treated as hyperparameters of the knowledge prediction model, and are tuned using the repeated cross-topic validation in the model fitting process (see Section 6.3.2). Some combinations of γ, β, τ which reduce the feature set to an empty set are excluded in the experiment.

Baseline. We compare the approach introduced in this chapter against the approach introduced in Chapter 5. In which we proposed to build classifiers using

[3] http://scikit-learn.org

user interaction features to predict KG and post-KS. The baseline model achieved best performance when using Random Forest as classifier and when applying certain thresholds on the feature-indicator-correlation and the between-feature-correlation. In the repeated cross validation process of our experiments in this chapter, we tune the hyperparameters of the baseline model again using grid search to ensure a fair comparison.

6.3.2 Evaluation Method

In order to estimate the performance and generalisation of pretrained and pretuned models on unseen topics, we conduct a repeated cross-topic validation consisting of an inner loop, where hyperparameters are tuned, and an outer loop, for the actual cross-topic performance assessment. Instead of randomly splitting the experimental dataset into training and testing set, we split the search sessions in our dataset according to the topic of a search session. More specifically, for the repeated cross validation, we run 11 iterations in the outer loop, for each run, we use the session data corresponding to one topic for testing, and the rest of the sessions for training and validation. Similar to the outer loop, the inner loop consists of 10 iterations, for each run, the session data corresponding to one topic is used for validation, the session data corresponding to the remaining 9 topics are used for model training. Hyperparameters of the classifiers as well as the feature selection thresholds α, β and τ are tuned in the inner loop. We evaluate the results according to the following metrics:

- *Accuracy (Accu) across all classes:* percentage of search sessions that were classified with the correct class label.
- *Precision (P), Recall (R), F1 (F1) score of class i:* we compute the standard precision, recall and F1 score on the prediction result of each class i.
- *Macro average of precision (P), recall (R), and F1 (F1):* the average of the corresponding score across 3 classes.

6.4 Results

Table 6.3 Best performing results of different approaches according to average F1 score.

KI	approach	feature cat.	classifier	low			moderate			high			average			Accu
				P	R	F1	P	R	F1	P	R	F1	P	R	F1	
pre-KS	new	WR&UB	rf	0.600	0.621	0.610	0.296	0.308	0.302	0.652	0.617	0.634	0.516	0.515	**0.515**	**0.549**
	baseline	-	rf	0.442	0.529	0.482	0.146	0.115	0.129	0.511	0.479	0.495	0.367	0.374	0.368	0.416
post-KS	new	WR&UB	nb	0.367	0.590	0.453	0.513	0.411	0.456	0.559	0.429	0.485	0.480	0.476	**0.465**	**0.464**
	baseline	-	rf	0.320	0.508	0.392	0.417	0.368	0.391	0.519	0.351	0.419	0.418	0.409	0.401	0.399
KG	new	WR&UB	lr	0.578	0.440	0.500	0.425	0.571	0.487	0.500	0.431	0.463	0.501	0.481	**0.483**	**0.485**
	baseline	-	rf	0.437	0.369	0.400	0.368	0.381	0.374	0.400	0.462	0.429	0.401	0.404	0.401	0.399

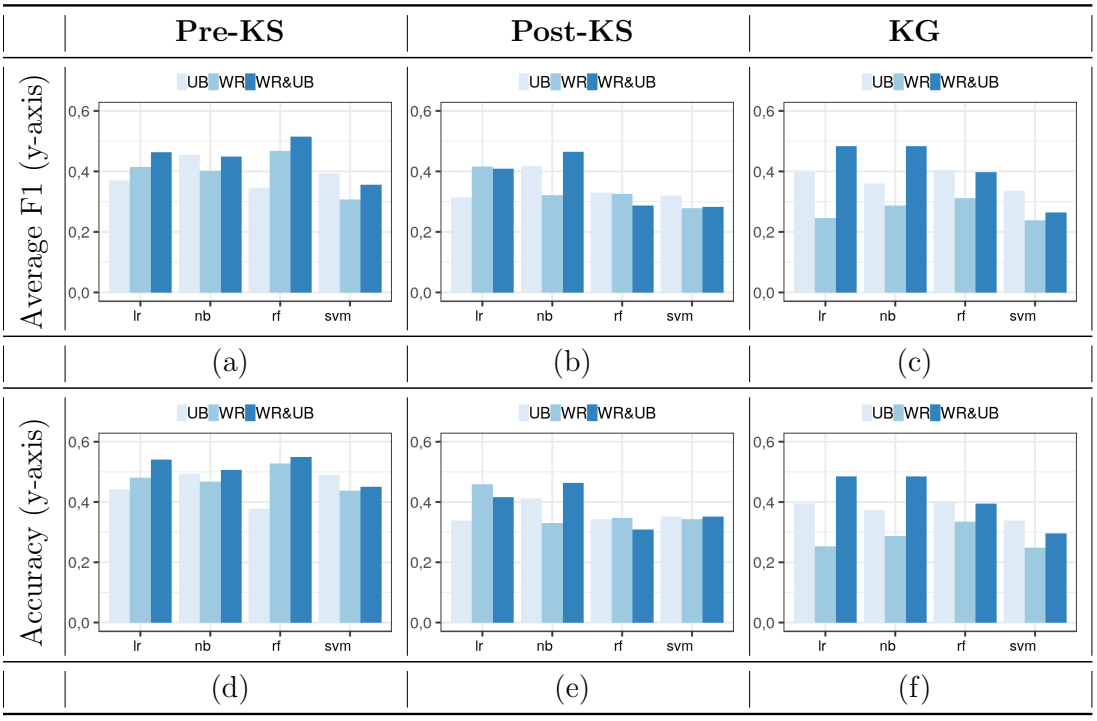


Figure 6.2 Average F1-score and accuracy for best performing classifier and respective feature category.

To evaluate the performance of our approach, we tune the hyperparameters according to the average F1 score through repeated cross-topic validation, as described in Section 6.3. We present the result of the configuration in terms of classifier and feature category that produces the highest average F1 score for each prediction task in Table 6.3. Our main findings are discussed below.

Overall performance. Using our approach, accuracy scores are above 0.464 for all 3 prediction tasks and the average F1 scores are above 0.465. Compared to the baseline, we observe improvements for all 3 prediction tasks, with the highest improvements for the pre-KS prediction task, where the average F1 score is 43.9% higher and the accuracy is 26.8% higher.

Knowledge indicator classes. Compared to the baseline, for pre-KS, our model shows particular improvements in F1 score for the moderate class, indicating that the resource features allow for better identifying medium knowledge state compared to the user behavior features. For post-KS our model shows similar improvements for all three classes. For knowledge gain, our model shows greater improvements for low and moderate KG classes.

The best performance with respect to both average F1-score and overall accuracy has been achieved for the pre-KS prediction, indicating that predicting the user's knowledge state on the search topic before the search session is a easier task compared to predicting the other two KIs. This is intuitive as the interactions such as input

queries and the resource selection are strongly affected by the user's pre-KS. While the post-KS is dependent on the pre-KS as well as the effort the user spends during the search session. Due to the short duration of the sessions in the ground truth dataset, despite using multiple features (e.g. *s_duration*, *b_time_avg_per_q*) to capture the effort of the user, it is more challenging to distinguish the post-KS and KG classes.

With respect to the prediction performance on different classes, we observe that for the pre-KS prediction, the model performs particularly well for low and high knowledge classes. For the prediction of post-KS and KG, on the other hand, performance differences on different KI classes are less pronounced.

In summary, our approach outperforms the baseline for all prediction tasks and the resource-related features appear to provide useful information for all the prediction tasks and knowledge classes. The performance of the classifiers using different categories of features and feature selection strategies will be discussed more in the remainder of this section.

6.4.1 Performance of Classifiers

Here we compare the performance achieved when using different classification algorithms, combined with the available feature categories, as seen in Figure 6.2. As also listed in Table 6.3, the best performing classifier varies for different prediction tasks. The *rf* classifier achieves the highest average F1-score for pre-KS prediction, outperforming the other classifiers by at least 11.3%. The *nb* classifer achieves the highest average F1-scores for the post-KS prediction task. The *lr* classifier achieves the highest average F1-score for the KG prediction, outperforming the *nb* classifier with a 0.1% score improvement. The result is inconsistent with the finding of previous work [YGH+18] where *rf* was the best performing classifier for both post-KS and KG prediction. The reasons behind might be: 1) different features, feature selection strategies and experimental setup (i.e. we test the models on unseen topics, we tune the feature selection thresholds as hyperparameters) and 2) the *rf* classifier used by previous work may have been overfitted. This is also supported by the intermediate results produced in the repeated cross-topic validation process, where we observed that the hyperparameters selected by the inner loop do not always produce the best average F1 score for the overall result. Hence, if the parameters are selected based on their performance on the test set, there is a high risk of overfitting. Both *lr* and *nb* are less prone to this effect which was observed to a larger extent for the two more challenging prediction problems.

6.4.2 Feature Category

The highest average F1 scores for all prediction tasks are achieved when using both Web resource and user behavior features. The results indicate that by utilizing signals from both categories, our approach is able to improve the performance of the

prediction models.

For pre-KS prediction using the *rf* classifier, using both categories of features only slightly outperforms using resource features only, with the average F1 score being 10.3% higher and the overall accuracy 4% higher. For post-KS prediction, using the *lr* classifier and resource features achieves similarly high accuracy compared to using *nb* and both categories of features. This suggests that the content of the Web resources a user interacted with might carry most of the meaningful signals for post-KS prediction. For KG prediction, none of the configurations using a single category of features achieves comparable results compared to the best performing configuration. The result suggests that both the user interaction and the visited resources have strong influence on user's knowledge gain and each group of features encodes unique information about the learning process.

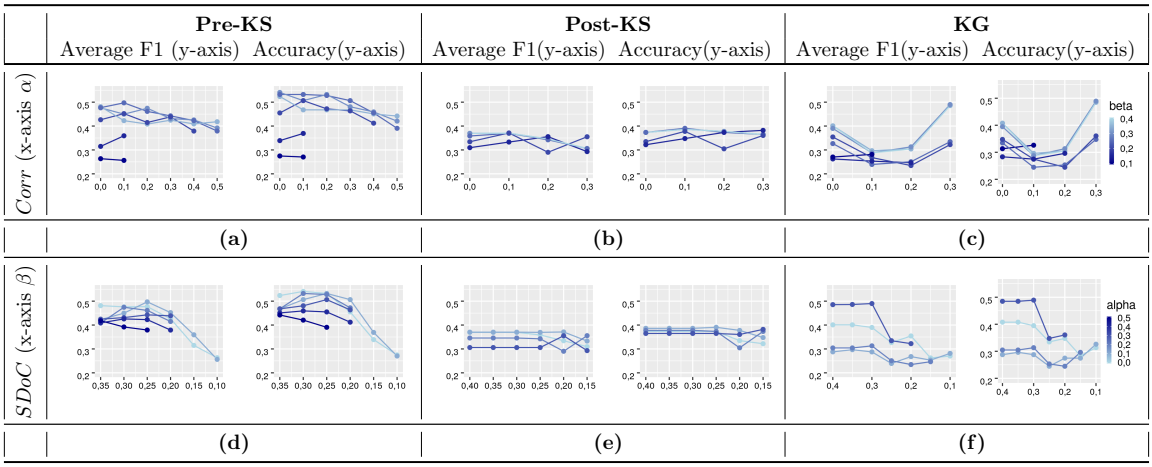


Figure 6.3 Classification performance of the different feature selection strategies using the complete set of features and the best performing classifiers for each of the prediction tasks (*rf* for Pre-KS, *nb* for Post-KS, and *lr* for KG). The threshold for the feature redundancy filter is fixed at $\tau = 0.9$.

6.4.3 Feature Selection Strategy

To better understand the interaction of feature selection strategies for the individual KIs, we evaluate the impact of settings for *feature effectiveness* ($Corr(f_i, KI) \geq \alpha$) and *generalizability* ($SDoC(f_i, KI) < \beta$) feature selection strategies on model performance. For each of the prediction tasks, we present the results of the best classification model using different feature selection configurations.

In Figure 6.3 (a), (b) and (c), the x-axis represents α, each line corresponds to a specific β, and vice versa for Figure 6.3 (d), (e) and (f). Larger values for α lead to fewer features while larger values for β lead to higher numbers of features – i.e. from left to right the filter settings are increasingly restrictive and darker colors show

more restrictive filter settings as well. The threshold for *feature redundancy* is fixed at $\tau = 0.9$, the most conservative value observed in the best performing classifier configurations.

In the pre-KS prediction task, low *feature effectiveness* thresholds of $\alpha \leq 0.2$ result in the best classification performance. More restrictive filter settings result in performance decreases for this prediction task. Similarly, the best performances are achieved with non-restrictive *generalizability* filter settings, i.e. $\beta \geq 0.25$. On their own, either of these filters removes useful features and results in a decrease in performance (both in terms of F1-score and accuracy); pairing $\beta = 0.35$ (does not remove any feature) with any $\alpha > 0$, for instance, results in a drop in F1-score from 0.481 to 0.425 or below. Nevertheless, a combination of moderate settings of $\alpha = 0.1$ and $\beta = 0.25$ selects 79 features (out of 136) that result in the best overall classification performance for this task: an F1-score of 0.497 (compared to 0.481 without filters) and Accuracy of 0.532 (compared to 0.524 without filters).

In the most challenging prediction problem, post-KS prediction, we observe a slightly positive impact in prediction accuracy when choosing a moderate *feature effectiveness* filter setting of $\alpha = 0.1$. A combination with the least restrictive *generalizability* filter setting that still removes features ($\beta = 0.25$) results in 57 features that allow the *nb* classifier to identify low and high knowledge classes better and improves its Accuracy from 0.373 to 0.391, while the average F1-score does not benefit due to a reduced recall for the medium class.

For the KG prediction task, there is overall a marked negative performance impact for introducing moderate *feature effectiveness* filter settings of $\alpha = 0.1$ and $\alpha = 0.2$, while the most restrictive setting of $\alpha = 0.3$ results in the highest performance, particularly when paired with the three least restrictive *generalizability* filter settings of $\beta \geq 0.3$. Within these settings, paired with $\alpha = 0.3$ there is no difference in the selected features, while more restrictive settings of $\beta < 0.3$ lead to a deterioration in performance. Applying the *feature effectiveness* filter in this prediction task improves F1-score from 0.401 to 0.490 and Accuracy from 0.408 to 0.489.

Overall, with regards to the *feature effectiveness* selection strategy (see Figure 6.3 (a), (b) and (c)), the best classification performance for each of the prediction tasks is achieved with $\alpha > 0$, confirming previous results that this is an effective strategy for reducing the feature set in our scenario. A similar observation can be made with respect to our additional *generalizability* selection strategy (see Figure 6.3 (d), (e) and (f)). Although for the filter settings with $\beta < 0.4$, the improvements are only minor and the effects of the filter vary across the different prediction tasks. In terms of the prediction tasks, the filters were least effective for the Post-KS prediction, which also showed the worst performance overall. In contrast, for KG prediction the *feature effectiveness* filter shows the largest effect, particularly for the logistic regression model.

6.5 Discussion

The experimental results suggest that it is possible to predict the user's knowledge state (change) without prior awareness of the specific learning intent of the user. Our approach outperforms SotA baselines on unseen topics by considering additional features of Web resources that users interact with.

However, while providing important contributions towards improving knowledge gain of users during Web search, the experimental results indicate that the current performance of predictive models requires improvement for real-world applications. Potential reasons for this might include (1) the limited scale of training data, (2) the lack of diversity of search sessions, in particular with respect to session length, and (3) issues related to our stratification approach when building classes for knowledge state (gain). Regarding (1), especially given that the topics in our experimental dataset are spanning across several different domains and considering the large number of features (179 features in total), the training data may not be sufficient for capturing the signals carried by all meaningful features. With respect to (2), the comparable short duration of all search sessions limits the signals provided for each feature. Certain features may provide more meaningful signals for longer search sessions only. Regarding (3), our stratification approach for separating knowledge state (change) classes using the *Standard Deviation Classification* approach may not be an ideal solution for user knowledge assessment. More targeted and domain-specific knowledge assessment methods may provide more meaningful classes, where classification performance may yield better results. Despite the aforementioned limitations, our experiments provide crucial insights into the effectiveness of a wide range of features and their use as part of supervised models for predicting knowledge gain and knowledge state of users during Web search.

Feature topic dependency. We conduct topic-dependency analysis on both Web resource features and user behavior features. Table 6.2 shows the user behavior features that are discussed in this section together with their $SDoC$ corresponding to each KI. More details about the complete list of user behavior features and their correlation to the KI can be found in Chapter 5. The 5 features with highest $SDoC(f_i, \textit{pre-KS}) \geq 0.292$ are *s_duration, s_duration_per_q, b_time_avg_per_q, b_revisited_ratio, q_uniq_term_ratio*. These user behavior features suggest that effort (e.g. session length) and browsing behavior are influenced by the topic itself and the knowledge gap of the user. Further, we observe that users are more likely to revisit pages during longer sessions on broad and complex topics.

The 5 features with highest $SDoC(f_i, \textit{post-KS}) \geq 0.283$ are *q_uniq_term_ratio, q_len_first, l_relig_avg, l_relativ_avg, q_uniq_term_first*. Unlike the result for pre-KS, more linguistic and query term related features are found to be topic dependent with respect to post-KS. A possible reason for this finding is that the different level of specificity of the topic might influence the observed words. Hence the assumption from previous work (e.g. [Arg14]) that higher knowledge state leads to higher coverage

of keywords in the query and resources may not hold for all topics.

The 5 features with highest $SDoC(f_i, KG) \geq 0.315$ are *b_revisited_ratio, m_scroll_dist, s_duration, l_percept_avg, m_rank_max*. Overall, the feature performance seems to vary more strongly for KG than for pre- and post-KS. Topic-dependency appears intuitive in a number of cases. For instance, in case of *l_percept_avg*, i.e. the number of perceptual process words (such as *see* or *hear*), may be specifically popular for certain topics. *l_bio_avg*, i.e. the number of biological process words, is an example of a highly domain-specific feature which contributes strongly to the overall performance in the pre-KS predication task. Our observations suggest that this feature contributes very strongly in life sciences-related topics, such as *Carpenter Bees*, *Altitude Sickness* or *HIV*. These findings underline that highly domain-specific linguistic features may provide very effective signals for KI prediction on unseen topics, in particular in more domain-specific models.

Correlation analysis. Whereas the correlation between user behavior features and KIs has been investigated in Chapter 4, here we focus on the Web resource features. We notice that the correlation between the Web resource features and different KIs varies strongly. For instance, the feature *c_verb_avg* is moderately positively correlated with pre- and post-KS and negatively correlated with KG.

The top 5 Web resource features positively correlated with pre-KS ($Corr(f_i, pre\text{-}KS) \geq 0.469$, $p < 0.05$) are *l_bio_avg, l_health_avg, l_focuspresent_avg, ls_you_avg and l_body_avg*. The top 5 features negatively correlated with pre-KS ($Corr(f_i, pre\text{-}KS) \leq -0.287$, $p < 0.05$) are *c_adj_avg, ls_article_avg, ls_Quote_avg, ls_number_avg, l_Analytic_avg*.

Similarly, for post-KS, the top 5 positively correlated ($Corr(f_i, pre\text{-}KS) \geq 0.302$, $p < 0.05$) resource features are *l_bio_avg, l_health_avg, l_focuspresent_avg, ls_you_avg, l_body_avg* and the top 5 negatively correlated ($Corr(f_i, pre\text{-}KS) \leq -0.294$, $p < 0.05$) resource features are *ls_article_avg, l_relativ_avg, h_oth_ul_avg, ls_Quote_avg, c_adj_avg*.

For KG, the top 5 positively correlated ($Corr(f_i, pre\text{-}KS) \geq 0.169$, $p < 0.05$) resource features are *l_Analytic_avg, ls_number_avg, c_noun_avg, l_anger_avg, l_money_avg* and the top 5 negatively correlated ($Corr(f_i, pre\text{-}KS) \leq -0.349$, $p < 0.05$) resource content features are *ls_Dic_avg, ls_conj_avg, l_focuspresent_avg, l_bio_avg, l_health_avg*. In particular with respect to the positive correlation, we observe that the amount of analytical words (*l_Analytic_avg*) correlates positively with KG. This is intuitively explained, assuming that analytical words may have higher occurrences in suitable learning material.

We observe that seemingly topic-dependent features such as the number of biological process words (*l_bio_avg*) correlate more strongly with the corresponding KI. This may be due to the selection of topics in the dataset we considered, which include a larger proportion of life sciences-related topics. Given that these features also proved useful in cross-topic prediction of KIs, we argue that sufficient coverage of domains may be desirable, as it may allow for capturing topic-dependent usefulness

of resources and thus improve domain-specific model performances even on unseen topics.

7 Conclusions and Future Work

In this thesis, we have identified and addressed some of the important problems in improving knowledge accessibility for both machines and humans. We proposed and evaluated several novel approaches to overcome the challenges in corresponding research fields. The KnowMore pipeline we introduced in Chapter 3 demonstrates the potential of using Web markup data for knowledge base augmentation. Our findings in Chapter 4 enrich the current understanding of search as learning, and our approaches (Chapter 5 and 6) have the potential to improve the search systems towards supporting human learning. In this chapter, we draw the main conclusions from our findings presented in this thesis and discuss directions for future work.

7.1 Conclusions

This thesis has addressed two main challenges, namely, (i) improving the accessibility of knowledge on the Web for machines through knowledge base augmentation using Web markup data, and (ii) improving the accessibility of knowledge on the Web for humans through understanding and supporting human learning in Web search.

In Chapter 3, we introduce *KnowMore*, an approach towards knowledge base augmentation from large-scale Web markup data, based on a combination of entity matching, data fusion and deduplication techniques. We apply our method to the WDC2015 corpus as the largest publicly available Web markup crawl (approx. 20 billion quads) and augment three established knowledge bases, namely Wikidata, Freebase and DBpedia. Evaluation results suggest superior performance of our approach with respect to novelty as well as correctness compared to state-of-the-art data fusion baselines, with an average F1 score increase across datasets of 0.142 (0.119) compared to the baseline *PrecRecCorr* (*CBFS*). Our experimental results indicate comparably consistent performance across a variety of types, whereas the performance of baseline methods tends to vary strongly. Our evaluation of the KBA task on two types demonstrates a strong potential to complement traditional knowledge bases

through data sourced from Web markup. We achieve a 100% coverage for particular properties, while providing substantial contributions to others. In addition, we demonstrate the capability to augment KBs with additional entity descriptions, particularly about long-tail entities, where for randomly selected entities of type *Product* from WDC, we are able to generate new entity descriptions with an average size of 6.45 facts.

Through our work presented in Chapter 4, we presented a study that investigates user knowledge gain through informational search sessions. We quantified the knowledge gain of users by calibrating their knowledge before they began a search session corresponding to a topic, and by assessing their knowledge after the session. We found a significant effect of information need on user queries and navigational patterns, but no direct effect on the knowledge gain. Users exhibited a higher knowledge gain through search sessions pertaining to topics they were less familiar with. Users who spent more active time on webpages depicted a higher knowledge gain. We also found a positive correlation between the average complexity of queries entered by users and their knowledge gain. Our findings revealed deeper insights into the search behavior of users in informational search sessions, and the impact of information needs on knowledge gain.

Chapter 5 introduces the supervised models we build for predicting user knowledge state and knowledge gain. The experimental results underline that a user's knowledge gain and knowledge state can be modeled based on a user's online interactions observable throughout the search process. Through feature analysis, we provide evidence for an improved understanding between individual user behavior and the corresponding knowledge state and change.

In Chapter 6, we propose to improve the performance and generalizability of the knowledge prediction models described in Chapter 5. We extracted a feature set, which extends our prior work with Web resource-centric features, and combine them with user behavior features. We also conducted a preliminary analysis with respect to the correlation and dependency of features to the KIs. For knowledge modeling, we applied and evaluated several feature selection strategies that focus on different aspects of feature effectiveness, showing that reducing the feature set and accounting for topic-dependency of features improves generalization performance. For each of the studied knowledge indicators, our approach outperforms the SotA baseline in the cross-topic experimental evaluation.

Alongside these results, we also make the gathered datasets available[1], [2], which provide resources that can facilitate further research in this area.

[1] http://l3s.de/~yu/knowmore/
[2] https://sites.google.com/view/predicting-user-knowledge/

7.2 Future Directions

While we made attempts to tackle the challenges in improving knowledge accessibility on the Web for machines and humans, there are still many issues that need to be addressed. Building on our observations and findings presented in this thesis, we plan to investigate the following aspects in the future.

Knowledge base augmentation. While our experiments have exploited the WDC corpus, we will consider more targeted Web crawls, which are better suited to augment entities (or properties) of a particular type or discipline. Here, targeted datasets which are retrieved with the dedicated aim to suit a particular KBA task are thought to further improve the KBA performance.

Another identified direction for future research is the investigation of the complementary nature of other sources of entity-centric Web data, for instance, data sourced from Web tables, when attempting to augment KBs.

Additional objectives for future work have surfaced during the experiments. For instance, identity resolution problems might occur during the matching step originating from different meanings of a particular entity. Current work aims at pre-clustering result sets into distinct entity meanings, from which we will be able to augment distinct disambiguated entity descriptions.

Finally, we are investigating an iterative approach which enables the generation of entity-centric knowledge graphs of a certain length (*hop-size*), rather than flat entity descriptions. This would further facilitate research into the generation of domain or type-specific knowledge graphs from distributed Web markup.

Understanding human learning in Web search. As a part of future work, we aim to reproduce and refine the findings in more varied search sessions, where durations and learning intents are more diverse; involving considerably longer search sessions and, for instance, procedural knowledge rather than intents focused on declarative knowledge only. This would provide the opportunity to observe evolution-oriented features, such as considering the evolution of queries, their length and complexity.

Potential applications of this work include the consideration of user knowledge and the expected learning progress of a user as part of Web search engines and information retrieval approaches, or within informal learning-oriented search settings, such as libraries or knowledge- and resource-centric online platforms.

User interface and interaction. Learning oriented online platforms (e.g. coursera[3], mooc[4], Didactalia[5]) have been constantly optimized to improve the learning performance of users. Examples are, for instance, the use of learning dashboards to inform users about his/her learning progress or provide discussion forums to enable collaboration among users. However, within general-purpose search engines, there is

[3] https://www.coursera.org/
[4] http://mooc.org/
[5] https://didactalia.net

a lack of attention for the support of learning, also due to the general-purpose nature of such environments and the variety of tasks conducted there. A central question for research in that area is whether and how interfaces can be adapted to improve learning performance even in such general-purpose environments. An attempt has been made by Arora [Aro15], by aiming at improving user engagement in learning oriented search tasks through providing a richer representation of retrieved Web documents. Specifically, they explored methods of finding useful semantic concepts within retrieved documents, with the objective of creating improved document surrogates for presentation in the SERP.

Retrieval and ranking. As current search engines are optimized by considering an information need disregarding the learning intent behind a query, relatively little research has been carried out on optimising retrieval and ranking algorithms towards particular learning needs. For instance, Dave et al. [DV+14] discussed the potential of two ranking models with varied objectives (i.e. paragraph retrieval model, dependency based re-ranking) on enhancing the performance of learning-centric search engines. Recently, Syed and Collins-Thompson [SCT17] proposed to optimize the learning outcome of the vocabulary learning task by selecting a set of documents while considering keyword density and domain knowledge of the learner. Their theoretical framework provides a sound basis for furthering the study on learning-oriented retrieval techniques.

Bibliography

[ABDR06] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10. ACM, 2006.

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *International Semantic Web Conference*, ISWC, pages 722–735, 2007.

[AGL13] Alessio Palmero Aprosio, Claudio Giuliano, and Alberto Lavelli. Extending the coverage of DBpedia properties using distant supervision over Wikipedia. In *Proceedings of the 2013th International Conference on NLP & DBpedia-Volume 1064*, pages 20–31. CEUR-WS. org, 2013.

[AKA+01] Lorin W Anderson, David R Krathwohl, P Airasian, K Cruikshank, R Mayer, P Pintrich, J Raths, and M Wittrock. A taxonomy for learning, teaching and assessing: A revision of bloom's taxonomy. *New York. Longman Publishing. Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction*, 9(2):137–175, 2001.

[Arg14] Jaime Arguello. Predicting search task difficulty. In *ECIR*, volume 14, pages 88–99, 2014.

[Aro15] Piyush Arora. Promoting user engagement and learning in amorphous search tasks. In *Proceedings of the 38th International ACM SIGIR Con-*

ference on Research and Development in Information Retrieval, pages 1051–1051. ACM, 2015.

[AWDB12] Eugene Agichtein, Ryen W. White, Susan T. Dumais, and Paul N. Bennet. Search, interrupted: Understanding and predicting search task continuation. In *SIGIR*. ACM, 2012.

[BB14] Volha Bryl and Christian Bizer. Learning conflict resolution strategies for cross-language Wikipedia data fusion. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1129–1134, Seoul, Korea, 2014. ACM New York, NY, USA. DOI: 10.1145/2567948.2578999.

[BBD+16] Jakob Beetz, Ina Blümel, Stefan Dietze, Besnik Fetahui, Ujwal Gadiraju, Martin Hecher, Thomas Krijnen, Michelle Lindlar, Martin Tamke, Raoul Wessel, et al. Enrichment and preservation of architectural knowledge. In *3D Research Challenges in Cultural Heritage II*, pages 231–255. Springer, 2016.

[BEP+08] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.

[BG19] Nilavra Bhattacharya and Jacek Gwizdka. Measuring learning during search: Differences in interactions, eye-gaze, and semantic similarity to expert knowledge. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 63–71. ACM, 2019.

[BL12] Lorenz Bühmann and Jens Lehmann. Universal OWL axiom enrichment for large knowledge bases. In Annette ten Teije, Johanna Völker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Acquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez, editors, *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, pages 57–71, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. DOI:10.1007/978-3-642-33876-2_8.

[BL13] Lorenz Bühmann and Jens Lehmann. Pattern based knowledge base enrichment. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web – ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 33–48,

Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-41335-3_3.

[BLHL+01] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[Blu12] Patrick Blumschein. Intentional learning. In *Encyclopedia of the Sciences of Learning*, pages 1600–1601. Springer, 2012.

[BN09] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1, 2009.

[Bro02] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[BS+89] Carl Bereiter, Marlene Scardamalia, et al. Intentional learning as a goal of instruction. *Knowing, learning, and instruction: Essays in honor of Robert Glaser*, pages 361–392, 1989.

[CES15] Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis. Entity resolution in the web of data. *Synthesis Lectures on the Semantic Web*, 5(3):1–122, 2015.

[CGL+13] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, 49(5):1075–1091, 2013.

[CMBJ97] Michael P Carey, Dianne Morrison-Beedy, and Blair T Johnson. The hiv-knowledge questionnaire: Development and evaluation of a reliable, valid, and practical self-administered questionnaire. *AIDS and Behavior*, 1(1):61–74, 1997.

[CTHH17] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. Search as Learning (Dagstuhl Seminar 17092). *Dagstuhl Reports*, 7(2):135–162, 2017.

[CTRHS16] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 163–172. ACM, 2016.

[DGH+14a] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM. DOI: 10.1145/2623330.2623623.

[DGH+14b] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. From data fusion to knowledge fusion. *Proc. VLDB Endow.*, 7(10):881–892, June 2014. DOI: 10.14778/2732951.2732962.

[DGM+15] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949, May 2015. DOI: 10.14778/2777598.2777603.

[Die98] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998. DOI: 10.1162/089976698300017197.

[DL07] Diana DeStefano and Jo-Anne LeFevre. Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, 23(3):1616–1641, 2007.

[DMS15] Arnab Dutta, Christian Meilicke, and Heiner Stuckenschmidt. Enriching structured knowledge with open information. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 267–277, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee. DOI: 10.1145/2736277.2741139.

[DTY+17] Stefan Dietze, Davide Taibi, Ran Yu, Phil Barker, and Mathieu d'Aquin. Analysing and improving embedded markup of learning resources on the Web. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 283–292, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. DOI: 10.1145/3041021.3054160.

[DV+14] Kushal Dave, Vasudeva Varma, et al. Computational advertising: Techniques for targeting relevant ads. *Foundations and Trends® in Information Retrieval*, 8(4–5):263–418, 2014.

[DYD19] Masoud Davari, Ran Yu, and Stefan Dietze. Understanding the influence of task difficulty on user fixation behavior. In *The 2nd International Workshop on ExplainAble Recommendation and Search (EARS), 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2019.

[ETWD14] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232. ACM, 2014.

[EW16] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48, 2016.

[FEMR15] Michael Färber, Basil Ell, Carsten Menne, and Achim Rettinger. A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web Journal*, 1(1):1–5, 2015.

[GC16] Jacek Gwizdka and Xueshu Chen. Towards observable indicators of learning on search. In *SAL@ SIGIR*, 2016.

[GHB+13] Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. Real-time RDF extraction from unstructured data streams. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web – ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 135–150, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-41335-3_9.

[GJG04] Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479. ACM, 2004.

[GS06] Jacek Gwizdka and Ian Spence. What can searching behavior tell us about the difficulty of information tasks? a study of web navigation. *Proceedings of the Association for Information Science and Technology*, 43(1):1–22, 2006.

[GYB17] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality–on the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 5–14. ACM, 2017.

[GYDH18] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing knowledge gain of users in informational search sessions on the web. In *2018 ACM on Conference on Human Information Interaction and Retrieval (CHIIR)*. ACM, 2018.

[HA17] Benjamin D Horne and Sibel Adali. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*, 2017.

[HCTCE07] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, 2007.

[HGBS13] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. From search session detection to search mission detection. In John P. McDermott, editor, *10th International Conference Open Research Areas in Information Retrieval (OAIR 13)*, pages 85–92. ACM, 5 2013.

[HHK+18] Anett Hoppe, Peter Holtz, Yvonne Kammerer, Ran Yu, Stefan Dietze, and Ralph Ewerth. Current challenges for studying search as learning processes. In *Workshop on Learning & Education with Web Data (LILE), 10th ACM Conference on Web Science (WebSci)*, 2018.

[Hol79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[HPV+16] Matthias Hagen, Martin Potthast, Michael Völske, Jakob Gomoll, and Benno Stein. How writers search: Analyzing the search and writing logs of non-fictional essays. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 193–202. ACM, 2016.

[JBS09] Bernard J Jansen, Danielle Booth, and Brian Smith. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, 45(6):643–663, 2009.

[JG11] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1148–1158, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[JK08] Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM*. ACM, 2008.

[KAEW15] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. Development and evaluation of search tasks for iir experiments using a

cognitive complexity framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 101–110. ACM, 2015.

[KBW+11] Alexander Kotov, Paul N. Bennett, Ryen W. White, Susan T. Dumais, and Jaime Teevan. Modeling and analysis of cross-session search tasks. In *SIGIR*. ACM, 2011.

[KFJRC75] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[KM12] Pallika H. Kanani and Andrew K. McCallum. Selecting actions for resource-bounded information extraction using reinforcement learning. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 253–262, New York, NY, USA, 2012. ACM. DOI: 10.1145/2124295.2124328.

[KMD+12] Tomáš Knap, Jan Michelfeit, Jakub Daniel, Petr Jerman, Dušan Rychnovskỳ, Tomáš Soukup, and Martin Nečaskỳ. ODCleanstore: A framework for managing and providing integrated Linked Data on the web. In X. Sean Wang, Isabel Cruz, Alex Delis, and Guangyan Huang, editors, *Web Information Systems Engineering - WISE 2012: 13th International Conference, Paphos, Cyprus, November 28-30, 2012. Proceedings*, pages 815–816, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-35063-4_74.

[Kri07] Klaus Krippendorff. Computing krippendorff's alpha reliability. *Departmental papers (ASC)*, page 43, 2007.

[KSGB12] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990, 2012.

[LABT11] Jens Lehmann, Sören Auer, Lorenz Bühmann, and Sebastian Tramp. Class expression learning for ontology engineering. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):71–81, 2011.

[LB08] Yuelin Li and Nicholas J Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, 2008.

[LH69] Mc Laughlin and G. Harry. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[MB12] Hannes Mühleisen and Christian Bizer. Web data commons-extracting structured data from two large web corpora. *LDOW*, 937:133–145, 2012.

[MBSJ09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Meu17] Robert Meusel. *Web-scale profiling of semantic annotations in HTML pages*. 2017.

[MK12] Jan Michelfeit and Tomáš Knap. Linked Data fusion in ODCleanstore. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track - Volume 914*, ISWC-PD'12, pages 45–48, Aachen, Germany, Germany, 2012. CEUR-WS.org.

[MMB12] Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: Linked Data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 116–123, New York, NY, USA, 2012. ACM. DOI: 10.1145/2320765.2320803.

[MP15] Robert Meusel and Heiko Paulheim. Heuristics for fixing common errors in deployed schema.org Microdata. In Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann, editors, *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 – June 4, 2015. Proceedings*, pages 152–168, Cham, 2015. Springer International Publishing. DOI: 10.1007/978-3-319-18818-8_10.

[MPB14] Robert Meusel, Petar Petrovski, and Christian Bizer. The WebDataCommons microdata, RDFa and microformat dataset series. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web – ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 277–292, Cham, 2014. Springer International Publishing. 10.1007/978-3-319-11964-9_18.

[MRP16] Robert Meusel, Dominique Ritze, and Heiko Paulheim. Towards more accurate statistical profiling of deployed schema.org Microdata. *J. Data and Information Quality*, 8(1):3:1–3:31, October 2016. DOI: 10.1145/2992788.

[PDSD+14] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 433–444, New York, NY, USA, 2014. ACM. DOI: 10.1145/2588555.2593674.

[PM+00] Dan Pelleg, Andrew W Moore, et al. X-means: Extending K-means with efficient estimation of the number of clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, 2000.

[PP13] Heiko Paulheim and Simone Paolo Ponzetto. Extending DBpedia with Wikipedia list pages. In *Proceedings of the 2013th International Conference on NLP & DBpedia - Volume 1064*, NLP-DBPEDIA'13, pages 85–90, Aachen, Germany, Germany, 2013. CEUR-WS.org.

[RL04] Daniel E Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM, 2004.

[RLB15] Dominique Ritze, Oliver Lehmberg, and Christian Bizer. Matching HTML tables to DBpedia. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, WIMS '15, pages 10:1–10:6, New York, NY, USA, 2015. ACM. DOI: 10.1145/2797115.2797118.

[RLOB16] Dominique Ritze, Oliver Lehmberg, Yaser Oulabi, and Christian Bizer. Profiling the potential of web tables for augmenting cross-domain knowledge bases. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 251–261, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. DOI: 10.1145/2872427.2883017.

[RM16] Petar Ristoski and Peter Mika. Enriching product ads with metadata from HTML annotations. In Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange, editors, *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 – June 2, 2016, Proceedings*, pages 151–167, Cham, 2016. Springer International Publishing. DOI: 10.1007/978-3-319-34129-3_10.

[SCMN13] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 926–934. Curran Associates, Inc., 2013.

[SCT17] Rohail Syed and Kevyn Collins-Thompson. Retrieval algorithms optimized for human learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–564. ACM, 2017.

[SCT18] Rohail Syed and Kevyn Collins-Thompson. Exploring document retrieval features associated with improved short-and long-term vocabulary learning outcomes. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*, pages 191–200. ACM, 2018.

[SGY+16] Pracheta Sahoo, Ujwal Gadiraju, Ran Yu, Sriparna Saha, and Stefan Dietze. Analysing structured scholarly data embedded in web pages. In Alejandra González-Beltrán, Francesco Osborne, and Silvio Peroni, editors, *Semantics, Analytics, Visualization. Enhancing Scholarly Data: Second International Workshop, SAVE-SD 2016, Montreal, QC, Canada, April 11, 2016, Revised Selected Papers*, pages 90–100, Cham, 2016. Springer International Publishing. DOI: 10.1007/978-3-319-53637-8_10.

[Sin12] Amit Singhal. Introducing the knowledge graph: things, not strings. *Official google blog*, 5, 2012.

[SKW07] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM. DOI: 10.1145/1242572.1242667.

[SMI+10] Andreas Schultz, Andrea Matteini, Robert Isele, Christian Bizer, and Christian Becker. LDIF - Linked Data integration framework. In *Proceedings of the Second International Conference on Consuming Linked Data - Volume 782*, COLD'11, pages 125–130, Aachen, Germany, Germany, 2010. CEUR-WS.org.

[Str87] A.L. Strauss. *Qualitative Analysis for Social Scientists*. Cambridge University Press, 1987.

[TFDCM16] Alberto Tonon, Victor Felder, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. VoldemortKG: Mapping schema.org and Web entities to Linked Open Data. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web – ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II*, pages 220–228, Cham, 2016. Springer International Publishing. DOI: 10.1007/978-3-319-46547-0_23.

[Vak16] Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.

[Vra12] Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*, pages 1063–1064. ACM, 2012.

[WDT09] Ryen W White, Susan T Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*, pages 132–141. ACM, 2009.

[WGM+14] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 515–526, New York, NY, USA, 2014. ACM. DOI: 10.1145/2566486.2568032.

[WKEA12] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. Grannies, tanning beds, tattoos and nascar: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 254–257. ACM, 2012.

[WT10] Gerhard Weikum and Martin Theobald. From information to knowledge: Harvesting entities and relationships from Web sources. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, pages 65–76, New York, NY, USA, 2010. ACM. DOI: 10.1145/1807085.1807097.

[YdG+19] Ran Yu, Mathieu d'Aquin, Dragan Gasevic, Joachim Kimmerle, Eelco Herder, and Ralph Ewerth. Lile2019: 8th international workshop on learning and education with web data. In *Companion Publication of the 10th ACM Conference on Web Science*, pages 15–16. ACM, 2019.

[YFGD16] Ran Yu, Besnik Fetahu, Ujwal Gadiraju, and Stefan Dietze. A survey on challenges in Web markup data for entity retrieval. In *International Semantic Web Conference (Posters & Demos), Kobe, Japan, October 17-21*, 2016.

[YGD18] Ran Yu, Ujwal Gadiraju, and Stefan Dietze. Detecting, understanding and supporting everyday learning in web search. In *Workshop on Learning & Education with Web Data (LILE), 10th ACM Conference on Web Science (WebSci)*, 2018.

[YGF+19a] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, Oliver Lehmberg, Dominique Ritze, and Stefan Dietze. Knowmore–knowledge base augmentation with structured web markup. *Semantic Web*, 1(10):159–180, 2019.

[YGF+19b] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, Oliver Lehmberg, Dominique Ritze, and Stefan Dietze. Knowmore–knowledge base augmentation with structured web markup. In *The Semantic Web – ISWC 2019 - Journal Track*, 2019.

[YGFD15] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, and Stefan Dietze. Adaptive focused crawling of linked data. In *International Conference on Web Information Systems Engineering*, pages 554–569. Springer, 2015.

[YGFD17] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, and Stefan Dietze. FuseM: Query-centric data fusion on structured Web markup. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 179–182. IEEE Computer Society, 2017. DOI: 10.1109/ICDE.2017.69.

[YGH+18] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting user knowledge gain in informational search sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 75–84. ACM, 2018.

[YGZ+16] Ran Yu, Ujwal Gadiraju, Xiaofei Zhu, Besnik Fetahu, and Stefan Dietze. Towards entity summarisation on structured web markup. In Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer, and Christoph Lange, editors, *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016, Revised Selected Papers*, pages 69–73, Cham, 2016. Springer International Publishing. DOI:10.1007/978-3-319-47602-5_15.

[YTR+19] Ran Yu, Rui Tang, Markus Rokicki, Ujwal Gadiraju, and Stefan Dietze. Topic-independent modeling of user knowledge in informational search sessions. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.

[ZCB11] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1225–1226. ACM, 2011.

[ZRGH12] Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han. A Bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, February 2012. DOI: 10.14778/2168651.2168656.

Appendix I

In Chapter 6, we have introduced 109 Web document-centric features and 70 user behavior-centric features for user knowledge state and knowledge gain prediction in informational search sessions. In order to improve the readability, we did not list all the features in Chapter 5 and 6. In Table A.1, we give the description of features and their metrics as described in Section 6.2.3. Prefixes of the feature names are selected based on their corresponding category, specifically:

- Web resource features:
  - *c_** denotes document complexity features.
  - *h_** denotes HTML structure related features.
  - *l_** and *ls_** denote linguistic features.
- User behavior features:
  - *b_** denotes user browsing behavior related features.
  - *m_** denotes features extracted from mouse movements.
  - *q_** denotes query term and query behavior-related features.
  - *SERP_** denotes features extracted from the user interaction on SERPs.

Table A.1 Considered Web resource features and user behavior features.

	Corr			SDoC			
Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
c_adj_avg	-0.287	-0.329	0.076	0.166	0.230	0.187	Ratio of the number of adjective to the total number of words
c_aoa_avg	0.265	0.199	-0.157	0.227	0.177	0.245	Average age-of-acquisition rating of words in each webpage \hline
c_char_avg	-0.077	-0.084	0.024	0.180	0.187	0.193	Average number of characters per term
c_fk_avg	-0.174	-0.203	0.043	0.236	0.155	0.284	Flesh-Kincaid Grad Readability Index
c_gi_avg	-0.062	0.092	0.153	0.257	0.272	0.235	Gunning Fog Grade Readability Index
c_noun_avg	-0.165	0.001	0.203	0.153	0.179	0.118	Ratio of the number of noun to the total number of words
c_oth_avg	0.117	-0.007	-0.149	0.143	0.150	0.105	Ratio of the number of other words to the total number of words
c_sentence_avg	-0.024	-0.006	0.024	0.230	0.181	0.257	Average number of words per sentence
c_smog_avg	-0.041	0.088	0.124	0.230	0.248	0.202	SMOG Readability Index
c_uniq_word_avg	0.015	0.164	0.118	0.202	0.178	0.229	Ratio of the number of unique words to the total number of words
c_verb_avg	0.342	0.234	-0.222	0.188	0.235	0.199	Ratio of the number of verb to the total number of words
c_word_avg	-0.191	-0.243	0.030	0.196	0.199	0.190	Number of words in each webpage
h_img_avg	-0.187	-0.273	0.001	0.204	0.191	0.191	Number of <img>elements
h_link_avg	-0.099	-0.240	-0.079	0.190	0.221	0.192	Number of outbound links
h_nav_ul_avg	0.206	0.184	-0.098	0.249	0.172	0.238	Number of elements embedded in<nav>elements
h_oth_ul_avg	-0.243	-0.305	0.043	0.218	0.148	0.187	Number of elements not in <nav>elements
h_p_avg	-0.281	-0.230	0.151	0.205	0.255	0.238	Average length of each paragraph in <p>elements
h_script_avg	0.165	0.067	-0.145	0.221	0.141	0.245	Number of <script>elements \hline
l_achieve_avg	0.058	-0.018	-0.087	0.195	0.184	0.276	Number of achievement words
l_affect_avg	0.067	0.006	-0.078	0.217	0.257	0.203	Nmber of affect words
l_affiliation_avg	0.278	0.200	-0.173	0.168	0.229	0.224	Number of affiliation words

	Corr			SDoC			
Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
l_Analytic_avg	-0.469	-0.294	0.329	0.192	0.224	0.172	Number of analytic words
l_anger_avg	-0.250	-0.151	0.179	0.172	0.210	0.090	Number of anger words
l_anx_avg	0.411	0.292	-0.259	0.136	0.147	0.165	Number of anxiety words
l_assent_avg	0.190	0.096	-0.152	0.265	0.145	0.306	Number of assent words
l_Authentic_avg	-0.146	-0.190	0.020	0.246	0.272	0.216	Nmber of authentic words
l_bio_avg	0.564	0.364	-0.386	0.238	0.207	0.281	Number of biological process words
l_body_avg	0.469	0.336	-0.293	0.236	0.186	0.264	Number of body words
l_cause_avg	0.317	0.117	-0.289	0.276	0.165	0.246	Number of causal words
l_certain_avg	0.299	0.078	-0.300	0.246	0.197	0.244	Number of certainty words
l_Clout_avg	0.370	0.286	-0.214	0.156	0.196	0.206	Nmber of clout words
l_cogproc_avg	0.408	0.180	-0.349	0.219	0.123	0.242	Number of cognitive process words
l_death_avg	-0.217	-0.128	0.159	0.198	0.197	0.239	Number of death words
l_differ_avg	0.454	0.287	-0.315	0.169	0.160	0.194	Number of differentiation words
l_discrep_avg	0.374	0.228	-0.268	0.194	0.225	0.159	Number of discrepancy words
l_drives_avg	0.174	0.096	-0.133	0.178	0.169	0.216	Number of core drives and needs
l_family_avg	0.247	0.144	-0.181	0.175	0.186	0.255	Number of family words
l_feel_avg	0.056	-0.040	-0.102	0.232	0.208	0.223	Number of feeling words
l_female_avg	0.197	0.140	-0.125	0.093	0.226	0.228	Numbe of female referents
l_filler_avg	0.052	-0.040	-0.097	nan	nan	nan	Number of filters
l_focusfuture_avg	0.184	0.163	-0.089	0.190	0.223	0.229	Number of future focus words
l_focuspast_avg	-0.258	-0.256	0.101	0.288	0.243	0.211	Number of past focus words
l_focuspresent_avg	0.514	0.302	-0.376	0.263	0.226	0.240	Number of present focus words
l_friend_avg	0.283	0.196	-0.182	0.200	0.200	0.223	Number of friend words
l_health_avg	0.556	0.351	-0.387	0.186	0.217	0.259	Number of health words
l_hear_avg	-0.115	-0.080	0.074	0.212	0.218	0.203	Numbe of hearing words
l_home_avg	-0.010	0.070	0.071	0.229	0.167	0.218	Number of home words
l_informal_avg	0.162	0.068	-0.141	0.178	0.153	0.141	Number of informal speech words
l_ingest_avg	0.170	0.215	-0.028	0.221	0.131	0.252	Number of ingesting words
l_insight_avg	0.176	0.012	-0.205	0.177	0.189	0.227	Number of insightful words
l_leisure_avg	0.045	0.074	0.008	0.257	0.214	0.232	Number of leisure words
l_male_avg	-0.058	-0.035	0.041	0.205	0.171	0.221	Numbe of male referents
l_money_avg	-0.156	-0.026	0.169	0.130	0.156	0.188	Number of money words

	Corr			SDoC			
Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
l_motion_avg	0.106	0.036	-0.099	0.197	0.194	0.200	Number of motion words
l_negemo_avg	-0.032	-0.035	0.010	0.184	0.233	0.149	Nmber of negative emotional words
l_netspeak_avg	0.146	0.063	-0.127	0.179	0.141	0.124	Number of netspeak words
l_nonflu_avg	0.195	0.129	-0.130	0.199	0.208	0.142	Number of nonfluencies
l_percept_avg	-0.043	-0.039	0.020	0.240	0.172	0.316	Number of perceptual processes
l_posemo_avg	0.197	0.061	-0.190	0.227	0.246	0.215	Nmber of positive emotion words
l_power_avg	-0.181	-0.118	0.123	0.194	0.164	0.219	Number of power words
l_relativ_avg	-0.264	-0.295	0.077	0.226	0.285	0.209	Number of relativity words
l_relig_avg	-0.121	-0.190	-0.011	0.160	0.292	0.248	Number of religion words
l_reward_avg	0.441	0.261	-0.321	0.248	0.155	0.217	Number of reward focus words
l_risk_avg	0.455	0.274	-0.328	0.205	0.238	0.186	Number of risk words
l_sad_avg	0.170	0.073	-0.146	0.238	0.221	0.269	Number of sadness words
l_see_avg	-0.077	-0.057	0.047	0.176	0.180	0.280	Number of seeing words
l_sexual_avg	0.444	0.233	-0.348	0.202	0.233	0.236	Number of sexual words
l_social_avg	0.436	0.298	-0.285	0.152	0.175	0.201	Number of social words
l_space_avg	-0.281	-0.262	0.125	0.188	0.269	0.279	Number of space words
l_swear_avg	0.082	-0.009	-0.108	0.213	0.206	0.103	Number of swear words
l_tentat_avg	0.431	0.259	-0.311	0.188	0.160	0.222	Number of tentative words
l_time_avg	-0.190	-0.243	0.029	0.187	0.216	0.136	Number of time words
l_Tone_avg	0.101	0.052	-0.080	0.212	0.264	0.162	Nmber of emotional tone words
l_work_avg	-0.060	-0.036	0.044	0.236	0.129	0.229	Number of work words
ls_adj_avg	0.174	-0.016	-0.226	0.170	0.196	0.188	Number of adjectives
ls_adverb_avg	0.363	0.131	-0.335	0.239	0.200	0.220	Number of common adverbs
ls_AllPunc_avg	-0.201	-0.200	0.079	0.234	0.184	0.256	Number of punctuation
ls_Apostro_avg	0.219	0.055	-0.222	0.200	0.200	0.163	Number of apostrophes
ls_article_avg	-0.297	-0.294	0.118	0.187	0.170	0.174	Number of articles
ls_auxverb_avg	0.415	0.224	-0.321	0.231	0.199	0.186	Number of auxiliary verbs
ls_Colon_avg	-0.032	-0.126	-0.066	0.260	0.146	0.223	Number of colons
ls_Comma_avg	-0.086	0.053	0.149	0.212	0.186	0.170	Number of commas
ls_compare_avg	0.260	0.096	-0.238	0.189	0.185	0.239	Number of comparatives
ls_conj_avg	0.422	0.184	-0.362	0.208	0.182	0.182	Number of conjunctions
ls_Dash_avg	-0.143	-0.240	-0.026	0.223	0.178	0.256	Number of dashes

	Corr			SDoC			
Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
ls_Dic_avg	0.398	0.164	-0.349	0.213	0.197	0.243	Number of dictionary words
ls_Exclam_avg	0.014	-0.003	-0.019	0.221	0.168	0.238	Number of exclamation marks
ls_function_avg	0.285	0.087	-0.277	0.183	0.183	0.156	Number of function words
ls_i_avg	0.369	0.185	-0.297	0.169	0.222	0.234	Number of I pronouns
ls_interrog_avg	0.400	0.238	-0.291	0.226	0.251	0.217	Number of interrogatives
ls_ipron_avg	0.282	0.082	-0.277	0.256	0.241	0.196	Number of impersonal pronouns
ls_negate_avg	0.412	0.258	-0.288	0.200	0.182	0.242	Number of negations
ls_number_avg	-0.358	-0.227	0.248	0.240	0.238	0.246	Number of numbers
ls_OtherP_avg	-0.162	-0.133	0.087	0.209	0.179	0.185	Number of other punctuation
ls_Parenth_avg	-0.158	-0.171	0.051	0.188	0.165	0.235	Number of parentheses (pairs)
ls_Period_avg	-0.121	-0.184	-0.005	0.213	0.217	0.241	Number of periods
ls_ppron_avg	0.427	0.283	-0.286	0.206	0.216	0.222	Number of personal pronouns
ls_prep_avg	0.124	-0.033	-0.179	0.173	0.171	0.240	Number of prepositions
ls_pronoun_avg	0.410	0.232	-0.308	0.207	0.255	0.167	Number of total pronouns
ls_QMark_avg	0.398	0.213	-0.309	0.163	0.197	0.208	Number of question marks
ls_quant_avg	0.279	0.152	-0.214	0.196	0.203	0.149	Number of quantifiers
ls_Quote_avg	-0.339	-0.320	0.147	0.208	0.158	0.260	Number of quotation marks
ls_SemiC_avg	0.014	-0.047	-0.056	0.271	0.238	0.257	Number of semicolor
ls_shehe_avg	-0.099	-0.120	0.020	0.192	0.086	0.191	Number of she or he pronouns
ls_Sixltr_avg	-0.052	0.078	0.129	0.221	0.192	0.170	Number of words which have more then 6 letters
ls_they_avg	0.004	0.050	0.037	0.163	0.214	0.097	Number of they pronouns
ls_verb_avg	0.430	0.217	-0.345	0.235	0.201	0.206	Number of regular verbs
ls_we_avg	0.152	0.081	-0.118	0.162	0.239	0.238	Number of we pronouns
ls_you_avg	0.473	0.337	-0.297	0.155	0.247	0.201	Number of you pronouns
b_distint_domain_num	0.166	0.042	-0.167	0.183	0.223	0.256	#domains
b_num	0.067	0.028	-0.059	0.284	0.187	0.297	# total pages
b_num_from_non_SERP	-0.066	-0.019	0.065	0.241	0.273	0.249	#document navigated from non-landing page
b_num_from_SERP	0.111	0.030	-0.110	0.178	0.223	0.235	#document navigated from landing page
b_num_per_q	-0.043	-0.038	0.020	0.249	0.209	0.256	# total pages / query

	Corr			SDoC			
Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
b_pct_from_non_SERP	-0.081	0.018	0.114	0.219	0.242	0.230	%document from non-landing page
b_pct_from_SERP	0.081	-0.018	-0.114	0.219	0.242	0.230	%document from landing page
b_revisited_ratio	0.002	-0.043	-0.038	0.292	0.180	0.357	Ratio of revisited pages
b_time_avg_per_page	-0.159	0.062	0.246	0.228	0.257	0.296	time active per query
b_time_avg_per_q	-0.144	0.035	0.205	0.306	0.257	0.264	average active time
b_time_max_per_page	-0.206	0.061	0.303	0.191	0.263	0.260	max active time
b_time_total	-0.114	0.061	0.191	0.260	0.240	0.285	total active time on documents
b_ttl_len_avg	0.397	0.281	-0.251	0.203	0.190	0.258	avg title len
b_ttl_len_max	0.347	0.239	-0.225	0.168	0.191	0.258	max title len
b_ttl_len_min	0.320	0.235	-0.195	0.247	0.222	0.183	min title len
b_ttl_len_total	0.198	0.106	-0.154	0.205	0.247	0.259	total title len
b_ttl_query_overlap_avg	-0.137	-0.024	0.147	0.149	0.202	0.187	avg query term appearance in title
b_ttl_query_overlap_max	-0.145	-0.011	0.168	0.138	0.188	0.155	max query term appearance in title
b_ttl_query_overlap_min	-0.062	-0.007	0.069				min query term appearance in title
b_ttl_query_overlap_total	-0.070	0.012	0.095	0.135	0.194	0.193	total query term appearance in title
b_uniq_num	0.085	0.064	-0.051	0.266	0.198	0.236	# unique pages
b_uniq_num_per_q	-0.043	-0.038	0.020	0.249	0.209	0.256	# unique pages / query
b_url_q_overlap_avg	-0.054	0.042	0.101	0.183	0.195	0.237	avg query term appearance in url
b_url_q_overlap_max	-0.092	0.029	0.137	0.137	0.239	0.203	max query term appearance in url
b_url_q_overlap_min	0.063	0.083	-0.008				min query term appearance in url
b_url_q_overlap_total	-0.023	0.053	0.072	0.133	0.242	0.186	total query term appearance in url
m_num	0.031	0.158	0.094	0.245	0.182	0.311	TotalMouseovers*: total number of mouseovers in the session.
m_num_per_q	-0.035	0.130	0.152	0.247	0.218	0.264	AvgMouseovers: average number of mouseovers per query.
m_rank_max	-0.027	0.110	0.126	0.194	0.187	0.315	MaxMouseover*: max mouseover rank in the session.
m_rank_max_per_q	-0.056	0.110	0.160	0.216	0.199	0.270	AvgMaxMouseover: average max mouseover rank per query.
m_scroll_dist	-0.030	-0.011	0.028	0.278	0.220	0.353	TotalScrollDistance*: total scroll distance in session.

	Corr			SDoC			
Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
m_scroll_dist_per_q	-0.044	-0.032	0.027	0.230	0.245	0.252	AvgScrollDistance: average scroll distance per query.
m_scroll_max_pos	-0.074	-0.009	0.083	0.169	0.221	0.313	MaxScrollPosition*: max scroll position in session.
m_scroll_max_pos_per_q	-0.076	-0.015	0.080	0.186	0.198	0.246	AvgMaxScrollPosition: average max scroll position per query.
q_complexity_amax_diff	0.036	0.031	-0.018				max complexity diff
q_complexity_avg	0.017	0.033	0.006	0.170	0.257	0.238	avg compexity
q_complexity_max	0.085	0.100	-0.021	0.162	0.223	0.215	max complexity
q_complexity_min	0.080	0.097	-0.017	0.166	0.209	0.133	min complexity
q_len_first	0.033	0.002	-0.039	0.180	0.298	0.197	first query length
q_len_las	0.152	0.080	-0.120	0.209	0.206	0.169	last query length
q_num	0.067	0.058	-0.033	0.255	0.225	0.217	#query
q_term_avg	0.122	0.076	-0.086	0.231	0.254	0.223	mean query length
q_term_max	0.085	0.057	-0.057	0.289	0.233	0.242	max query length
q_term_min	0.126	0.077	-0.090	0.135	0.195	0.191	min query length
q_term_total	0.050	0.044	-0.025	0.281	0.164	0.235	total query length
q_uniq_term_avg	0.123	0.087	-0.078	0.206	0.215	0.173	avg uniq terms per query
q_uniq_term_first	0.023	0.019	-0.012	0.179	0.283	0.184	first query uniq terms
q_uniq_term_last	0.173	0.094	-0.133	0.186	0.216	0.165	last query uniq terms
q_uniq_term_max	0.090	0.079	-0.044	0.283	0.216	0.212	max uniq terms
q_uniq_term_min	0.132	0.082	-0.093	0.139	0.186	0.197	min uniq terms
q_uniq_term_ratio	-0.185	-0.085	0.154	0.292	0.304	0.234	uniq term ratio
q_uniq_term_total	0.051	0.046	-0.024	0.284	0.213	0.228	#uniq terms
s_duration	0.113	0.087	-0.066	0.380	0.262	0.351	session duration
s_duration_per_q	0.113	0.087	-0.065	0.312	0.234	0.288	session_len_per_query
SERP_avg_time_to_first_click	-0.034	-0.055	-0.004	0.287	0.229	0.295	avg time to first click
SERP_click	0.119	0.037	-0.115	0.148	0.274	0.189	total click
SERP_click_interval	0.118	0.012	-0.134	0.148	0.200	0.178	average click per query
SERP_click_per_query	0.026	-0.042	-0.067	0.166	0.267	0.151	average click interval
SERP_click_rank_avg	0.071	0.024	-0.067	0.286	0.189	0.280	mean ranking
SERP_click_rank_highest	-0.013	0.090	0.091				highest ranking

	Corr			SDoC			
Feature Name	Pre-KS	Post-KS	KG	Pre-KS	Post-KS	KG	Feature Description
SERP_click_rank_lowest	0.107	-0.012	-0.141	0.199	0.184	0.196	lowest ranking
SERP_no_click_query_num	0.046	0.037	-0.026	0.264	0.248	0.231	queries with no click
SERP_no_click_query_pct	0.071	0.016	-0.073	0.265	0.217	0.234	ratio of queries with no click
SERP_time_avg	-0.028	0.010	0.043	0.264	0.268	0.255	average time active on SERPs
SERP_time_max	0.016	-0.007	-0.026	0.229	0.250	0.199	max time active on SERPs
SERP_time_total	0.052	0.023	-0.045	0.262	0.220	0.222	total active time on SERPs