Semantic Matching in Search

Hang Li
Huawei Technologies, Hong Kong
hangli.hl@huawei.com

Jun Xu
Huawei Technologies, Hong Kong
nkxujun@gmail.com

the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

H. Li and J. Xu. *Semantic Matching in Search*. Foundations and Trends® in Information Retrieval, vol. 7, no. 5, pp. 343–469, 2013.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-805-8

Foundations and Trends® in Information Retrieval
Volume 7, Issue 5, 2013

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2013, Volume 7, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in Information Retrieval
Vol. 7, No. 5 (2013) 343–469

DOI: 10.1561/1500000035

Semantic Matching in Search

Hang Li
Huawei Technologies, Hong Kong
hangli.hl@huawei.com

Jun Xu
Huawei Technologies, Hong Kong
nkxujun@gmail.com

Contents

Abstract

Relevance is the most important factor to assure users' satisfaction in search and the success of a search engine heavily depends on its performance on relevance. It has been observed that most of the dissatisfaction cases in relevance are due to term mismatch between queries and documents (e.g., query "ny times" does not match well with a document only containing "New York Times"), because term matching, i.e., the bag-of-words approach, still functions as the main mechanism of modern search engines. It is not exaggerated to say, therefore, that mismatch between query and document poses the most critical challenge in search. Ideally, one would like to see query and document match with each other, if they are topically relevant. Recently, researchers have expended significant effort to address the problem. The major approach is to conduct semantic matching, i.e., to perform more query and document understanding to represent the meanings of them, and perform better matching between the enriched query and document representations. With the availability of large amounts of log data and advanced machine learning techniques, this becomes more feasible and significant progress has been made recently. This survey gives a systematic and detailed introduction to newly developed machine learning technologies for query document matching (semantic matching) in search, particularly web search. It focuses on the fundamental problems, as well as the state-of-the-art solutions of query document matching on form aspect, phrase aspect, word sense aspect, topic aspect, and structure aspect. The ideas and solutions explained may motivate industrial practitioners to turn the research results into products. The methods introduced and the discussions made may also stimulate academic researchers to find new research directions and approaches. Matching between query and document is not limited to search and similar problems can be found in question answering, online advertising, cross-language information retrieval, machine translation, recommender systems, link prediction, image annotation, drug design, and other applications, as the general task of matching between objects from two different spaces. The technologies

introduced can be generalized into more general machine learning techniques, which is referred to as learning to match in this survey.

H. Li and J. Xu. *Semantic Matching in Search*. Foundations and Trends® in Information Retrieval, vol. 7, no. 5, pp. 343–469, 2013.

DOI: 10.1561/1500000035.

1

Introduction

1.1 Query Document Mismatch

A successful search engine must be good at relevance, coverage, freshness, response time, and user interface. Among them, relevance [156, 171, 157] is the most important factor, which is also the focus of this survey.

This survey mainly takes general web search as example. The issues discussed are not limited to web search, however; they exist in all the other searches such as enterprise search, desktop search, as well as question answering.

Search still heavily relies on the bag-of-words approach or term based approach. That is, queries and documents are represented as bags of words (terms), documents are indexed based on document terms, 'relevant' documents are retrieved based on query terms, the relevance scores between queries and retrieved documents are calculated on the basis of matching degrees between query terms and document terms, and finally the retrieved documents are ranked based on the relevance scores. This simple approach works quite well in practice and it still forms the foundation of modern search systems [131, 52, 6].

Table 1.1: Examples of query document mismatch.

query	document	term match	semantic match
seattle best hotel	seattle best hotels	partial	yes
pool schedule	swimming pool schedule	partial	yes
natural logarithm transform	logarithm transform	partial	yes
china kong	china hong kong	partial	no
why are windows so expensive	why are macs so expensive	partial	no

The bag-of-words approach also has limitations, however. It sometimes suffers from the query document mismatch drawback. For the majority of the cases of dissatisfaction reported at a commercial web search engine, in which users complain they cannot find information while the information does exist in the system, the reasons are due to mismatch between queries and documents. Similar trends are observed in other studies (cf., [206, 207])

A high matching degree at term level does not necessarily mean high relevance, and vice versa. For example, if the query is "ny times" and the document only contains "New York Times", then the matching degree of the query and the document at term level is low, although they are relevant. More examples of query document mismatch are given in Table 1.1.[1]

Query document mismatch occurs, when the searcher and author use different terms (representations) to describe the same concept, and this phenomenon is prevalent due to the nature of human language, i.e., the same meaning can be represented by different expressions and the same expression can represent different meanings. According to Furnas et al., on average 80-90% of the times, two people will name the same concept with different representations [67].

[1] China Kong is an American actor.

Table 1.2: Queries about "distance between sun and earth".

"how far" earth sun	average distance from the earth to the sun
"how far" sun	how far away is the sun from earth
average distance earth sun	average distance from earth to sun
how far from earth to sun	distance from earth to the sun
distance from sun to earth	distance between earth and the sun
distance between earth & sun	distance between earth and sun
how far earth is from the sun	distance from the earth to the sun
distance between earth sun	distance from the sun to the earth
distance of earth from sun	distance from the sun to earth
"how far" sun earth	how far away is the sun from the earth
how far earth from sun	distance between sun and earth
how far from earth is the sun	how far from the earth to the sun
distance from sun to the earth	

Table 1.2 shows example queries representing the same search need "distance between sun and earth" and Table 1.3 shows example queries representing the same search need "Youtube", collected from the search log of a commercial search engine [117]. Ideally, we would like to see the search system return the same results for the different variants of the queries. Web search engines, however, still cannot effectively satisfy the requirement. This is another side of the mismatch problem.

In web search, query document mismatch more easily occurs on tail pages and tail queries. This is because for head pages and head queries, usually there is more information attached to them. A head page may have a large number of anchor texts and associated queries in search log and they provide with the page different representations. The matching degree will be high, if the query matches with any of the representations. This seldom happens to a tail page, however. Mismatch, thus, is a typical example of the long tail challenge in search.

1.2 Semantic Matching in Search

The fundamental reason for mismatch is that no language analysis is conducted in search. Language understanding by computer is hard,

Table 1.3: Queries about "Youtube".

yutube	yuotube	yuo tube
ytube	youtubr	yu tube
youtubo	youtuber	youtubecom
youtube om	youtube music videos	youtube videos
youtube	youtube com	youtube co
youtub com	you tube music videos	yout tube
youtub	you tube com yourtube	your tube
you tube	you tub	you tube video clips
you tube videos	www you tube com	wwww youtube com
www youtube	www youtube com	www youtube co
yotube	www you tube	www utube com
ww youtube com	www utube	www u tube
utube videos	utube com	utube
u tube com	utub	u tube videos
u tube	my tube	toutube
outube	our tube	toutube

however, if not impossible. A more realistic approach beyond bag-of-words, referred to as semantic matching in this survey, would be to conduct more query analysis and document analysis to represent the meanings of the query and the document with richer representations and then perform query document matching with the representations. The analysis may include term normalization, phrase analysis, word sense analysis, topic analysis, and structure analysis, and the matching may be performed on form aspect, phrase aspect, word sense aspect, topic aspect, and structure aspect, as shown in Figure 1.1. Intuitively, if the meanings of the query and the document represented by the aspects are the same, then they should match each other well and thus be regarded relevant. In practice, the more aspects of the query and document can match, the more likely the query and document are relevant. With semantic matching, we can expect that the query document mismatch challenge can be successfully conquered.

Term normalization, including word segmentation for Asian languages, compounding for European languages, spelling error correction for European languages, should usually be carried out before query document matching. We refer to term normalization as matching on the form aspect. Query document matching on the phrase aspect means that the two should match at phrase level, not word level. For example, if the query is “hot dog”, then it should be recognized as a phrase and match the exactly same phrase in the document, but should not separately match words “hot” and “dog” in the document. Matching on the word sense aspect is to have phrases in the query and the document having the same sense match each other. For example, “ny” should match “New York”. If the query and the document have the same topics, then they should match on the topic aspect. For example, if the query is “microsoft office” and the document is about Microsoft Word, PowerPoint, and Excel, then the two should match in terms of topic. Query and document can also match on the structure aspect, where structure means linguistic structure. For example, the query “distance between sun and earth” matches with the document title “how far is sun from earth” (note that the two expressions have very different linguistic structures).

We can also consider query document matching on other aspects, for example, semantic class and named entity. We will discuss this in Section 9 on conclusion and open problems.

Semantic matching is also a term used in other fields in computer science, which represents a notion different from this survey. Given two graph-like structures, e.g., two database schemas, semantic matching is defined as an operator that identifies the nodes in the two structures which semantically correspond to each other [73].

Semantic matching also differs from the so-called semantic search, which has different definitions by different researchers. One of them is aimed at enriching search results of a conventional search system, by using information from semantic web (e.g., [77]). For example, the search result of query “yo-yo ma” is augmented by the cellist’s image, concert schedule, music albums, etc. in the semantic search. The semantic search by Bast et al. asks the user to formulate a

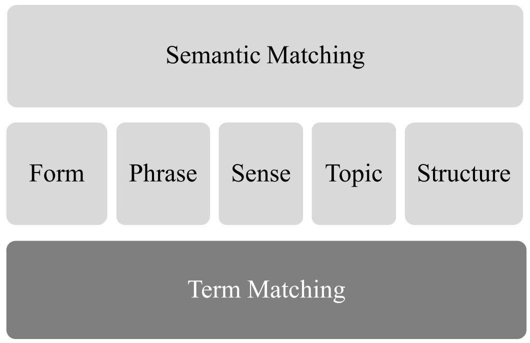

Figure 1.1: Semantic matching: if the meanings of the query and document represented in the aspects of form, phrase, sense, topic, and structure are the same, then they should match each other and be regarded relevant.

query with operators describing relations between entities, combines the information found from both documents and ontology, and returns to the user. Special search needs such as "finding plants with edible leaves and native to Europe" are supported [11]. In contrast, the semantic matching which we are concerned with here is carried out inside the search engine and users do not need to do anything different from conventional search.

Figure 1.2 illustrates the difference between semantic matching and semantic search. Semantic matching is concerned with search of documents by query, where both documents and query are unstructured data. Semantic search is usually concerned with search of documents and knowledge base by query, where documents and query are unstructured data, but knowledge base is structured data.

Query document mismatch has been studied in the long history of information retrieval (IR). In traditional IR, methods such as query expansion, pseudo-relevance feedback, and latent semantic indexing (LSI) have been intensively investigated and widely utilized. Nowadays large amounts of log data have been collected in web search and advanced machine learning techniques have been developed. We can really leverage big data and machine learning to more effectively

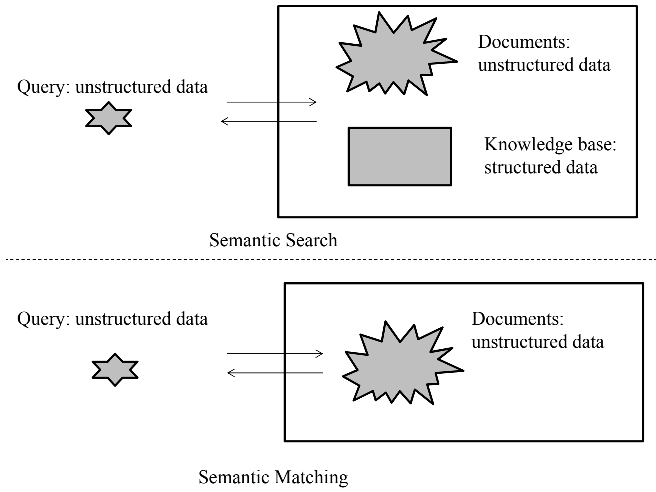

Figure 1.2: Semantic matching versus semantic search.

address the challenge of query document mismatch, as explained in this survey.

1.3 Matching and Ranking

In traditional IR, the distinction between ranking and matching in search is not made clear. Given a query, documents are retrieved from the index and matching between the query and each of the documents is carried out. The relevance of the document with respect to the query is represented as the matching degree between the two, calculated using an IR model (matching model) such as BM25 or language models for information retrieval (LM4IR). After that, the documents are ranked (sorted) based on their matching scores. In such a framework, matching scores and ranking scores are equivalent. [2]

Things have changed in web search. Importance of documents (web pages) is found useful for relevance ranking, and importance scores of

[2]We note that in modern web search not only relevance but also freshness, diversity, and other factors are considered. We restrict ourselves to relevance in this survey.

web pages calculated by models such as PageRank need to be incorporated into the ranking mechanism. Besides, many signals indicating the relevance (matching) degrees between queries and documents are also available and matching scores representing the signals can be calculated. How to combine the matching scores and importance scores then becomes a critical question. A simple approach is to linearly combine the scores and manually tune the weights. More sophisticated machine learning techniques for automatically constructing the ranking model using training data can also be considered. In fact, machine learning techniques for the purpose, referred to as learning to rank, have been intensively studied and widely applied in web search [128, 115]. Thus, in web search, the processes of matching and ranking are logically separated (first matching and then ranking).

As explained below, machine learning techniques for learning matching degrees between queries and documents (in general, heterogeneous objects), which are referred to as learning to match in this paper, have been developed. Learning to match is in fact *feature learning* for learning to rank, from the viewpoint of machine learning.

1.4 Semantic Matching in Other Tasks

Other tasks in information retrieval and natural language processing also rely on *semantic matching*, such as paraphrasing & textual entailment [62, 54], question answering [21], cross-language information retrieval (CLIR) [141, 140], online advertising [31], similar document detection [32, 33], and short text conversation [176, 130]. Table 1.4 summarizes the characteristics of the tasks.

For instance, CLIR is a subfield of information retrieval concerning with the problem of receiving queries in one language while retrieving documents in another language. Translation of either query or document from one language to another is naturally required in the task. Mismatch between query and document in two languages poses an even greater challenge to CLIR and matching on form aspect (compounding, word segmentation, spelling error correction), sense aspect (selection of translation), and topic aspect has also been tried and verified to be helpful [141, 140].

For another instance, online advertising makes use of web to deliver marketing messages and attract consumers. It usually involves publishers, who display advertisements at their web sites, and advertisers, who provide advertisements. Given some advertisements, it is necessary to find appropriate web sites for displaying them, i.e. conduct effective matching between publishers' content and advertisers' advertisements. Mismatch is also inevitable here. Methods have been proposed for addressing the mismatch challenge at form aspect, sense aspect, and topic aspect [31].

Short text conversation is a research problem proposed recently [176, 130]. It consists of one round of conversation between human and computer, with the former being a message from human and the latter being a comment on the message from the computer. Short text conversation constitutes one step of natural language conversation, and it also subsumes question answering as special case. Semantic matching between messages and comments needs also be considered, in a retrieval based approach in which a large collection of message and comment pairs is indexed, and given a message the most appropriate comment is retrieved, selected, and returned. Methods have been proposed to address the mismatch problem in the task as well [176, 130].

1.5 Machine Learning for Semantic Matching in Search

A natural question arises whether it is possible to use machine learning techniques to automatically learn the models for semantic matching in search. This is exactly the problem we address in this survey.

The task can be formalized as learning of matching model $f(q, d)$ or conditional probability model $P(r|q, d)$ using supervised learning techniques or learning of conditional probability model $P(q|d)$ using unsupervised learning techniques, where q denotes query, d denotes document, and r denotes relevance level. Note that here query and document are regarded as different (heterogeneous) objects.

Table 1.4: Characteristics of tasks that need semantic matching. Two natural language texts (A and B) are involved in the tasks.

task	types of texts	relation between texts
search	A=query, B=document	relevance
question answering	A=question, B=answer	answer to question
cross-language IR	A=query, B=document (in diff. lang.)	relevance
short text conversation	A=text, B=text	message and comment
similar document detection	A=text, B=text	similarity
online advertising	A=query, B=ads.	relevance as ads.
paraphrasing	A=sentence, B=sentence	equivalence
textual entailment	A=sentence, B=sentence	entailment

Different models can be defined, explicitly or implicitly representing semantic matching, i.e., matching on different aspects such as form aspect, phrase aspect, sense aspect, topic aspect, and structure aspect. Since query document mismatch is a long tail phenomenon, it is necessary to assume that no single signal is enough and construct matching models on different aspects and combine the uses of them in relevance ranking.

The following are some well-studied approaches, including matching by query reformulation, matching with term dependency model, matching with translation model, matching with topic model, and matching with latent space model. This survey will explain the approaches in detail.

Matching by query reformulation aims at reformulating the query so that it can have a better match with the semantically equivalent expressions in the documents. Reformulation of query includes spelling error correction, word splitting, word merging, and so on. The major issues with regard to query reformulation include re-writing of the original query, blending of the search results by the original query and reformulated queries, mining of similar queries, as well as query expansion.

A straightforward extension of the bag-of-words approach would be to perform matching based on multiple words in the query and document. This is exactly the process depicted in the term dependency models. One can represent different matching relations between the query terms and the document terms with the models, for example, co-occurrence of terms in both the query and document. Intuitively, if several terms co-occur within both the query and document, then they may represent the same concept and indicate stronger relevance.

Matching between the query and a part of the document, for example, the title, can be modeled as paraphrasing or translation in which a language expression is transformed into another language expression. Taking matching as a statistical translation task has been proposed previously and the approach has made significant progress in web search recently, in part because a large amount of click-through data becomes available and can be utilized as training data.

Given a collection of documents, topic modeling techniques can help find the topics of the documents, in which each topic is represented by a number of words. Probabilistic and non-probabilistic models have been proposed. In search, the topics of the query and the topics of the documents can be detected, and matching between the query and documents can be carried out with the topics.

We can represent queries and documents in two different vector spaces, map them into a hidden semantic space with lower dimensionality on the basis of query document associations in click-through data, and conduct matching between queries and documents in the latent semantic space. This is the basic idea of the approach of matching with latent space models. Many traditional IR models such as vector space model (VSM), BM25, and LM4IR can be interpreted as special cases of the latent space models, and thus the latent space models are quite fundamental for IR.

Matching between two heterogenous objects is not limited to search. It exists in many other applications, including paraphrasing & textual entailment, question answering, online advertising, cross-language information retrieval, similar document detection, short text conversation, machine translation, recommender systems (collaborative filtering), link prediction, image annotation, and drug design. It is necessary and important to generalize the techniques developed in different applications to a more general machine learning methodology in order to study the techniques more deeply and broadly. We refer to it as learning to match in this survey.

1.6 About This Survey

This survey focuses on the fundamental problems, as well as the state-of-the-art solutions of query document matching in search. The ideas and solutions explained may motivate industrial practitioners to turn the research results into products. The methods introduced and the discussions made may also stimulate academic researchers to find new research directions and approaches.

Section 2 gives a formulation of machine learning for query document matching in search and shows an implementation of it in web search. Sections 3-7 describe the five groups of learning techniques for query document matching, namely matching by query reformulation, matching with term dependency model, matching with translation model, matching with topic model, and matching with latent space model. Section 8 describes generalization of the techniques, learning to match, and introduce methods for collaborative filtering and paraphrasing & textual entailment. Section 9 summarizes the survey and discusses open problems. Sections 2-8 are self-contained, and thus the reader can choose sections to read on the basis of their interest and need.

This survey focuses more on machine learning and semantic matching. Several survey papers or books cover some related topics, such as LM4IR [204], query expansion [40], search and browse log mining [163, 94], and feature centric view on IR [135]. The reader is also encouraged to refer to the materials.

We assume that the reader has certain knowledge on machine learning and information retrieval. Those who want to know more about the fundamentals of the areas should refer to related books and papers. The machine learning techniques concerned with in this survey include statistical language model [204], statistical machine translation [99], learning to rank [128, 115, 116], graphical model [24], topic model [25], matrix factorization [103], kernel methods [158], sparse methods [3], and deep learning [4]. Explanations on the basic techniques in information retrieval can be found in the text books on IR [131, 52, 6].

[3] A tutorial on sparse methods by Bach can be found at www.di.ens.fr/fbach/.
[4] Tutorials on deep learning can be found at www.deeplearning.net/tutorial/.

References

[1] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *J. Mach. Learn. Res.*, 10:803–826, June 2009.

[2] Deepak Agarwal and Bee-Chung Chen. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 19–28, New York, NY, USA, 2009. ACM.

[3] Farooq Ahmad and Grzegorz Kondrak. Learning a spelling error model from search query logs. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 955–962, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[4] Jaime Arguello, Jonathan L. Elsas, Jamie Callan, and Jaime G. Carbonell. Document representation and query expansion models for blog recommendation. In *International Conference on Weblogs and Social Media*, pages 10–18. AAAI Press, 2008.

[5] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query clustering for boosting web page ranking. In JesÍšs Favela, Ernestina Menasalvas, and Edgar ChÍćvez, editors, *Advances in Web Intelligence*, volume 3034 of *Lecture Notes in Computer Science*, pages 164–175. Springer Berlin Heidelberg, 2004.

[6] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search.* Pearson Education Ltd., Harlow, England, 2nd edition, 2011.

[7] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Supervised semantic indexing. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 187–196, New York, NY, USA, 2009. ACM.

[8] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Learning to rank with (a lot of) word features. *Inf. Retr.*, 13(3):291–314, June 2010.

[9] Suhrid Balakrishnan and Sumit Chopra. Collaborative ranking. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 143–152, New York, NY, USA, 2012. ACM.

[10] Niranjan Balasubramanian, Giridhar Kumaran, and Vitor R. Carvalho. Exploring reductions for long web queries. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 571–578, New York, NY, USA, 2010. ACM.

[11] Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haussmann. Broccoli: Semantic full-text search at your fingertips. *CoRR*, 2012.

[12] Doug Beeferman and Adam Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '00, pages 407–416, New York, NY, USA, 2000. ACM.

[13] Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David Grossman, David D. Lewis, Abdur Chowdhury, and Aleksandr Kolcz. Automatic web query classification using labeled and unlabeled training data. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 581–582, New York, NY, USA, 2005. ACM.

[14] Michael Bendersky and W. Bruce Croft. Analysis of long queries in a large scale search log. In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 8–14, New York, NY, USA, 2009. ACM.

[15] Michael Bendersky and W. Bruce Croft. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 941–950, New York, NY, USA, 2012. ACM.

[16] Michael Bendersky, W. Bruce Croft, and David A. Smith. Joint annotation of search queries. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL '11, pages 102–111. The Association for Computer Linguistics, 2011.

[17] Michael Bendersky, Donald Metzler, and W. Bruce Croft. Learning concept importance using a weighted dependence model. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 31–40, New York, NY, USA, 2010. ACM.

[18] Michael Bendersky, Donald Metzler, and W. Bruce Croft. Effective query formulation with multiple information sources. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 443–452, New York, NY, USA, 2012. ACM.

[19] Yoshua Bengio, Holger Schwenk, Jean-SÍębastien SenÍęcal, FrÍęderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In DawnE. Holmes and LakhmiC. Jain, editors, *Innovations in Machine Learning*, volume 194 of *Studies in Fuzziness and Soft Computing*, pages 137–186. Springer Berlin Heidelberg, 2006.

[20] Paul N. Bennett, Krysta Svore, and Susan T. Dumais. Classification-enhanced ranking. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 111–120, New York, NY, USA, 2010. ACM.

[21] Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 192–199, New York, NY, USA, 2000. ACM.

[22] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 222–229, New York, NY, USA, 1999. ACM.

[23] Shane Bergsma and Qin Iris Wang. Learning noun phrase query segmentation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 819–826, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[24] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[25] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.

[26] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[27] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. The query-flow graph: Model and applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 609–618, New York, NY, USA, 2008. ACM.

[28] Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 211–218, New York, NY, USA, 2002. ACM.

[29] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 286–293, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[30] Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Lance Riedel, and Jeffrey Yuan. Online expansion of rare queries for sponsored search. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 511–520, New York, NY, USA, 2009. ACM.

[31] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM.

[32] Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, pages 21–, Washington, DC, USA, 1997. IEEE Computer Society.

[33] Andrei Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, COM '00, pages 1–10, London, UK, UK, 2000. Springer-Verlag.

[34] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311, June 1993.

[35] Fan Bu, Hang Li, and Xiaoyan Zhu. String re-writing kernel. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 449–458, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[36] Fan Bu, Hang Li, and Xiaoyan Zhu. An introduction to string re-writing kernel. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI'13, pages 2982–2986. AAAI Press, 2013.

[37] Robin D Burke, Kristian J Hammond, Vladimir Kulyukin, Steven L Lytinen, Noriko Tomuro, and Scott Schoenberg. Question answering from frequently asked question files: Experiences with the faq finder system. *AI magazine*, 18(2):57, 1997.

[38] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 243–250, New York, NY, USA, 2008. ACM.

[39] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 875–883, New York, NY, USA, 2008. ACM.

[40] Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, January 2012.

[41] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the world-wide web. *Comput. Netw. ISDN Syst.*, 27(6):1065–1073, April 1995.

[42] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 380–388, New York, NY, USA, 2002. ACM.

[43] Qing Chen, Mu Li, and Ming Zhou. Improving query spelling correction using web search results. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 181–189. ACL, 2007.

[44] TianQi Chen, Hang Li, Qiang Yang 0001, and Yong Yu. General functional matrix factorization using gradient boosting. In *ICML '13: Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *JMLR Proceedings*, pages 436–444, 2013.

[45] Tianqi Chen, Zhao Zheng, Qiuxia Lu, Weinan Zhang, and Yong Yu. Feature-based matrix factorization. *CoRR*, abs/1109.2271, 2011.

[46] David Chiang. Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228, June 2007.

[47] Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. Latent semantic analysis for text segmentation. In Lillian Lee and Donna Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, EMNLP '01, pages 109–117, 2001.

[48] C.W. Cleverdon. *The Effect of Variations in Relevance Assessment in Comparative Experimental Tests of Index Languages*. Cranfield Library report. Cranfield Inst. of Technology, 1970.

[49] Michael Collins and Nigel Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 263–270, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[50] Nick Craswell and Martin Szummer. Random walks on the click graph. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 239–246, New York, NY, USA, 2007. ACM.

[51] W. Bruce Croft, Michael Bendersky, Hang Li, and Gu Xu. Query representation and understanding workshop. *SIGIR Forum*, 44(2):48–53, January 2011.

[52] W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[53] Silviu Cucerzan and Eric Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, EMNLP '04, pages 293–300. ACL, 2004.

[54] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190. Springer-Verlag, Berlin, Heidelberg, 2006.

[55] Dipanjan Das and Noah A. Smith. Paraphrase identification as probabilistic quasi-synchronous recognition. 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 1 - Volume 1*, ACL '09, pages 468–476, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[56] Ali Dasdan, Paolo D'Alberto, Santanu Kolay, and Chris Drome. Automatic retrieval of similar content using search engine query interface. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 701–710, New York, NY, USA, 2009. ACM.

[57] Fabio De Bona, Stefan Riezler, Keith Hall, Massimiliano Ciaramita, Amaç Herdağdelen, and Maria Holmqvist. Learning dense models of query similarity from user click logs. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 474–482, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[58] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[59] Fernando Diaz. Regularizing ad hoc retrieval scores. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 672–679, New York, NY, USA, 2005. ACM.

[60] Fernando Diaz and Donald Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 154–161, New York, NY, USA, 2006. ACM.

[61] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drugĺctarget interactions: a brief review. *Briefings in Bioinformatics*, page bbt056, 2013.

[62] Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[63] Huizhong Duan and Bo-June (Paul) Hsu. Online spelling correction for query completion. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 117–126, New York, NY, USA, 2011. ACM.

[64] Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21, 1997.

[65] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.*, 29(2):8:1–8:34, April 2011.

[66] Edward A. Fox and Joeseph A. Shaw. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, volume 500-215 of *NIST Special Publication*, pages 243–252. NIST, 1994.

[67] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, November 1987.

[68] Jianfeng Gao, Xiaodong He, and Jian-Yun Nie. Clickthrough-based translation models for web search: from word models to phrase models. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1139–1148, New York, NY, USA, 2010. ACM.

[69] Jianfeng Gao and Jian-Yun Nie. Towards concept-based translation models using search logs for query expansion. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1:1–1:10, New York, NY, USA, 2012. ACM.

[70] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 170–177, New York, NY, USA, 2004. ACM.

[71] Jianfeng Gao, Kristina Toutanova, and Wen-tau Yih. Clickthrough-based latent semantic models for web search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 675–684, New York, NY, USA, 2011. ACM.

[72] Jianfeng Gao, Wei Yuan, Xiao Li, Kefeng Deng, and Jian-Yun Nie. Smoothing clickthrough data for web search ranking. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 355–362, New York, NY, USA, 2009. ACM.

[73] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. S-match: an algorithm and an implementation of semantic matching. In *Proceedings of The Semantic Web: Research and Applications, First European Semantic Web Symposium*, ESWS '04, pages 61–75. Springer, 2004.

[74] Andrew R. Golding and Dan Roth. A winnow-based approach to context-sensitive spelling correction. *Mach. Learn.*, 34(1-3):107–130, February 1999.

[75] Jagadeesh Gorla, Stephen Robertson, Jun Wang, and Tamas Jambor. A theory of information matching. *CoRR*, abs/1205.5569, 2012.

[76] David Grangier and Samy Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1371–1384, August 2008.

[77] R. Guha, Rob McCool, and Eric Miller. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 700–709, New York, NY, USA, 2003. ACM.

[78] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 267–274, New York, NY, USA, 2009. ACM.

[79] Jiafeng Guo, Gu Xu, Hang Li, and Xueqi Cheng. A unified and discriminative model for query refinement. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 379–386, New York, NY, USA, 2008. ACM.

[80] Matthias Hagen, Martin Potthast, Anna Beyer, and Benno Stein. Towards optimum query segmentation: In doubt without. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1015–1024, New York, NY, USA, 2012. ACM.

[81] Matthias Hagen, Martin Potthast, Benno Stein, and Christof Bräutigam. Query segmentation revisited. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 97–106, New York, NY, USA, 2011. ACM.

[82] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 362–370, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[83] David R. Hardoon and John Shawe-taylor. Kcca for different level precision in content-based image retrieval. In *In Third International Workshop on Content-Based Multimedia Indexing, IRISA*, 2003.

[84] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, December 2004.

[85] Michael Heilman and Noah A. Smith. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 1011–1019, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[86] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *In Proceedings of the International Conference on Articial Neural Networks*, pages 97–102, 1999.

[87] Dustin Hillard, Stefan Schroedl, Eren Manavoglu, Hema Raghavan, and Chirs Leggetter. Improving ad relevance in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 361–370, New York, NY, USA, 2010. ACM.

[88] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[89] Chien-Kang Huang, Lee-Feng Chien, and Yen-Jen Oyang. Relevant term suggestion in interactive web search based on contextual information in query session logs. *J. Am. Soc. Inf. Sci. Technol.*, 54(7):638–649, May 2003.

[90] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 2333–2338, New York, NY, USA, 2013. ACM.

[91] Samuel Huston, J. Shane Culpepper, and W. Bruce Croft. Indexing word sequences for ranked retrieval. *ACM Trans. Inf. Syst.*, 32(1):3:1–3:26, January 2014.

[92] Aminul Islam and Diana Inkpen. Real-word spelling correction using google web it 3-grams. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1241–1249, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[93] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[94] Daxin Jiang, Jian Pei, and Hang Li. Mining search and browse logs for web search: A survey. *ACM Trans. Intell. Syst. Technol.*, 4(4):57:1–57:37, October 2013.

[95] Rong Jin, Alex G. Hauptmann, and Cheng Xiang Zhai. Title language model for information retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 42–48, New York, NY, USA, 2002. ACM.

[96] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 387–396, New York, NY, USA, 2006. ACM.

[97] In-Ho Kang and GilChang Kim. Query type classification for web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 64–71, New York, NY, USA, 2003. ACM.

[98] Maryam Karimzadehgan and ChengXiang Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 323–330, New York, NY, USA, 2010. ACM.

[99] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[100] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[101] Yehuda Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 426–434, New York, NY, USA, 2008. ACM.

[102] Yehuda Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Trans. Knowl. Discov. Data*, 4(1):1:1–1:24, January 2010.

[103] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[104] Alexander Kotov and ChengXiang Zhai. Tapping into knowledge base for concept feedback: Leveraging conceptnet to improve search results for difficult queries. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 403–412, New York, NY, USA, 2012. ACM.

[105] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 61–68, New York, NY, USA, 2009. ACM.

[106] Oren Kurland and Lillian Lee. Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 194–201, New York, NY, USA, 2004. ACM.

[107] T. K. Landauer, D. Laham, and M. Derr. From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5214–5219, apr 2004.

[108] Hao Lang, Donald Metzler, Bin Wang, and Jin-Tao Li. Improved latent concept expansion using hierarchical markov random fields. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 249–258, New York, NY, USA, 2010. ACM.

[109] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.

[110] Matthew Lease. An improved markov random field model for supporting verbose queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 476–483, New York, NY, USA, 2009. ACM.

[111] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

[112] Joon Ho Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 267–276, New York, NY, USA, 1997. ACM.

[113] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 391–400, New York, NY, USA, 2005. ACM.

[114] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-based techniques for cross-language information retrieval. *Inf. Process. Manage.*, 41(3):523–547, May 2005.

[115] Hang Li. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 4(1):1–113, 2011.

[116] Hang Li. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94-D(10):1854–1862, 2011.

[117] Hang Li, Gu Xu, W. Bruce Croft, Michael Bendersky, Ziqi Wang, and Evelyne Viegas. Qru-1: A public dataset for promoting query representation and understanding research. In *Proceedings of the Workshop on Web Search Click Data*, WSCD '12, 2012.

[118] Mu Li, Yang Zhang, Muhua Zhu, and Ming Zhou. Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 1025–1032, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[119] Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 339–346, New York, NY, USA, 2008. ACM.

[120] Yanen Li, Huizhong Duan, and ChengXiang Zhai. A generalized hidden markov model with discriminative training for query spelling correction. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 611–620, New York, NY, USA, 2012. ACM.

[121] Yanen Li, Bo-Jun Paul Hsu, ChengXiang Zhai, and Kuansan Wang. Unsupervised query segmentation using clickthrough for information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 285–294, New York, NY, USA, 2011. ACM.

[122] Yinghao Li, Wing Pong Robert Luk, Kei Shiu Edward Ho, and Fu Lai Korris Chung. Improving weak ad-hoc queries using wikipedia as external corpus. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 797–798, New York, NY, USA, 2007. ACM.

[123] Zhen Liao, Daxin Jiang, Enhong Chen, Jian Pei, Huanhuan Cao, and Hang Li. Mining concept sequences from large-scale search logs for context-aware query suggestion. *ACM Trans. Intell. Syst. Technol.*, 3(1):17:1–17:40, October 2011.

[124] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, May 2007.

[125] Dekang Lin and Patrick Pantel. Discovery of inference rules for question-answering. *Nat. Lang. Eng.*, 7(4):343–360, December 2001.

[126] Kenneth C Litkowski. Question-answering using semantic relation triples. In *In Proceedings of the 8th Text Retrieval Conference (TREC-8)*, pages 349–356, 1999.

[127] H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, October 2004.

[128] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, March 2009.

[129] Yumao Lu, Fuchun Peng, Gilad Mishne, Xing Wei, and Benoit Dumoulin. Improving web search relevance with semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 648–657, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[130] Zhengdong Lu and Hang Li. A deep architecture for matching short texts. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1367–1375. Curran Associates, Inc., 2013.

[131] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[132] K. Tamsin Maxwell and W. Bruce Croft. Compact query term selection using topically related text. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 583–592, New York, NY, USA, 2013. ACM.

[133] Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 469–478, New York, NY, USA, 2008. ACM.

[134] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECML PKDD'11, pages 437–452, Berlin, Heidelberg, 2011. Springer-Verlag.

[135] Donald Metzler. *A Feature-Centric View of Information Retrieval*. Springer, 2012 edition, 2011.

[136] Donald Metzler and W. Bruce Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM.

[137] Donald Metzler and W. Bruce Croft. Latent concept expansion using markov random fields. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 311–318, New York, NY, USA, 2007. ACM.

[138] Alessandro Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, pages 318–329, Berlin, Heidelberg, 2006. Springer-Verlag.

[139] Alessandro Moschitti and Fabio Massimo Zanzotto. Fast and effective kernels for relational learning from texts. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 649–656, New York, NY, USA, 2007. ACM.

[140] Jian-Yun Nie. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125, 2010.

[141] Douglas W Oard and Anne R Diekema. Cross-language information retrieval. *Annual Review of Information Science (ARIST)*, 33, 1998.

[142] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 295–302, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[143] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[144] Deepa Paranjpe. Learning document aboutness from implicit user feedback and document structure. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 365–374, New York, NY, USA, 2009. ACM.

[145] Jae Hyun Park, W. Bruce Croft, and David A. Smith. A quasi-synchronous dependence model for information retrieval. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 17–26, New York, NY, USA, 2011. ACM.

[146] Fuchun Peng, Nawaaz Ahmed, Xin Li, and Yumao Lu. Context sensitive stemming for web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 639–646, New York, NY, USA, 2007. ACM.

[147] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 160–169, New York, NY, USA, 1993. ACM.

[148] Steffen Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, IEEE Computer Society, 2010.

[149] Steffen Rendle. Factorization machines with libfm. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, May 2012.

[150] Stefan Riezler and Yi Liu. Query rewriting using monolingual statistical machine translation. *Comput. Linguist.*, 36(3):569–582, September 2010.

[151] Eric Sven Ristad and Peter N. Yianilos. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):522–532, May 1998.

[152] S. E. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[153] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[154] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *Proceedings of the 2005 international conference on Subspace, Latent Structure and Feature Selection*, SLSFS'05, pages 34–51, Berlin, Heidelberg, 2006. Springer-Verlag.

[155] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.

[156] Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[157] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):2126–2144, November 2007.

[158] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

[159] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S. Davis. Human detection using partial least squares analysis. In *IEEE 12th International Conference on Computer Vision*, pages 24–31. IEEE, 2009.

[160] Daniel Sheldon, Milad Shokouhi, Martin Szummer, and Nick Craswell. Lambdamerge: Merging the results of query reformulations. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 795–804, New York, NY, USA, 2011. ACM.

[161] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Query enrichment for web-query classification. *ACM Trans. Inf. Syst.*, 24(3):320–352, July 2006.

[162] Lixin Shi and Jian-Yun Nie. Using various term dependencies according to their utilities. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1493–1496, New York, NY, USA, 2010. ACM.

[163] Fabrizio Silvestri. Mining query logs: Turning search usage data into knowledge. *Found. Trends Inf. Retr.*, 4(1—2):1–174, January 2010.

[164] Amit Singhal and Fernando Pereira. Document expansion for speech retrieval. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 34–41, New York, NY, USA, 1999. ACM.

[165] Richard Socher, Eric H. Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y. Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 801–809. Curran Associates, Inc., 2011.

[166] Ruihua Song, Michael J. Taylor, Ji-Rong Wen, Hsiao-Wuen Hon, and Yong Yu. Viewing term proximity from a different perspective. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 346–357. Springer-Verlag, Berlin, Heidelberg, 2008.

[167] Krysta M. Svore, Pallika H. Kanani, and Nazan Khan. How good is a span of terms?: exploiting proximity to improve web retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 154–161, New York, NY, USA, 2010. ACM.

[168] Bin Tan and Fuchun Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 347–356, New York, NY, USA, 2008. ACM.

[169] Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. Language model information retrieval with document expansion. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 407–414, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[170] Tao Tao and ChengXiang Zhai. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 295–302, New York, NY, USA, 2007. ACM.

[171] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. How users assess web pages for information seeking. *J. Am. Soc. Inf. Sci. Technol.*, 56(4):327–344, February 2005.

[172] Kristina Toutanova and Robert C. Moore. Pronunciation modeling for improved spelling correction. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 144–151, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[173] Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, UK, UK, 2001. Springer-Verlag.

[174] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[175] Chen Wang, Keping Bi, Yunhua Hu, Hang Li, and Guihong Cao. Extracting search-focused key n-grams for relevance ranking in web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 343–352, New York, NY, USA, 2012. ACM.

[176] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 935–945. ACL, 2013.

[177] Jianqiang Wang and Douglas W. Oard. Matching meaning for cross-language information retrieval. *Inf. Process. Manage.*, 48(4):631–653, July 2012.

[178] Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 187–194, New York, NY, USA, 2009. ACM.

[179] Quan Wang, Zheng Cao, Jun Xu, and Hang Li. Group matrix factorization for scalable topic modeling. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 375–384, New York, NY, USA, 2012. ACM.

[180] Quan Wang, Jun Xu, Hang Li, and Nick Craswell. Regularized latent semantic indexing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 685–694, New York, NY, USA, 2011. ACM.

[181] Quan Wang, Jun Xu, Hang Li, and Nick Craswell. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Trans. Inf. Syst.*, 31(1):5:1–5:44, January 2013.

[182] Xuanhui Wang and ChengXiang Zhai. Mining term association patterns from search logs for effective query reformulation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 479–488, New York, NY, USA, 2008. ACM.

[183] Ziqi Wang, Gu Xu, Hang Li, and Ming Zhang. A fast and accurate method for approximate string search. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 52–61, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[184] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

[185] Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Clustering user queries of a search engine. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 162–168, New York, NY, USA, 2001. ACM.

[186] Haocheng Wu, Yunhua Hu, Hang Li, and Enhong Chen. Query segmentation for relevance ranking in web search. *CoRR*, abs/1312.0182, 2013.

[187] Wei Wu, Hang Li, and Jun Xu. Learning query and document similarities from click-through bipartite graph with metadata. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 687–696, New York, NY, USA, 2013. ACM.

[188] Wei Wu, Zhengdong Lu, and Hang Li. Learning bilinear model for matching queries and documents. *J. Mach. Learn. Res.*, 14(1):2519–2548, January 2013.

[189] Wei Wu, Jun Xu, Hang Li, and Satoshi Oyama. Learning a robust relevance model for search using kernel methods. *J. Mach. Learn. Res.*, 12:1429–1458, July 2011.

[190] Gu Xu, Shuang-Hong Yang, and Hang Li. Named entity mining from click-through data using weakly supervised latent dirichlet allocation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1365–1374, New York, NY, USA, 2009. ACM.

[191] Jingfang Xu and Gu Xu. Learning similarity function for rare queries. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 615–624, New York, NY, USA, 2011. ACM.

[192] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA, 1996. ACM.

[193] Jun Xu, Hang Li, and Chaoliang Zhong. Relevance ranking using kernels. In Pu-Jen Cheng, Min-Yen Kan, Wai Lam, and Preslav Nakov, editors, *Information Retrieval Technology*, volume 6458 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2010.

[194] Jun Xu, Wei Wu, Hang Li, and Gu Xu. A kernel approach to addressing term mismatch. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 153–154, New York, NY, USA, 2011. ACM.

[195] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 267–273, New York, NY, USA, 2003. ACM.

[196] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. Optimizing web search using web click-through data. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 118–126, New York, NY, USA, 2004. ACM.

[197] Xiaobing Xue, Yu Tao, Daxin Jiang, and Hang Li. Automatically mining question reformulation patterns from search log data. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 187–192, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[198] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 523–530, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[199] Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 34–43, New York, NY, USA, 2009. ACM.

[200] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 29–41, Berlin, Heidelberg, 2009. Springer-Verlag.

[201] Xing Yi and James Allan. Discovering missing click-through query language information for web search. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 153–162, New York, NY, USA, 2011. ACM.

[202] Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 247–256, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[203] Fabio massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. A machine learning approach to textual entailment recognition. *Nat. Lang. Eng.*, 15(4):551–582, October 2009.

[204] ChengXiang Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2(3):137–213, March 2008.

[205] Xian Zhang, Yu Hao, Xiaoyan Zhu, Ming Li, and David R. Cheriton. Information distance from a question to an answer. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 874–883, New York, NY, USA, 2007. ACM.

[206] Le Zhao and Jamie Callan. Term necessity prediction. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 259–268, New York, NY, USA, 2010. ACM.

[207] Le Zhao and Jamie Callan. Automatic term mismatch diagnosis for selective query expansion. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 515–524, New York, NY, USA, 2012. ACM.