Advertising Keyword Suggestion Using Relevance-Based Language Models from Wikipedia Rich Articles

Amir H. Jadidinejad*, Fariborz Mahmoudi

Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Received 23 January 2011; accepted 10 March 2013

Abstract

When emerging technologies such as Search Engine Marketing (SEM) face tasks that require human level intelligence, it is inevitable to use the knowledge repositories to endow the machine with the breadth of knowledge available to humans. Keyword suggestion for search engine advertising is an important problem for sponsored search and SEM that requires a goldmine repository of knowledge. A recent strategy in this area is bidding on non-obvious yet relevant keywords, which are economically more viable. In this paper, we exploited a modified relevance-based language model for keyword suggestion problem using Wikipedia as our knowledge base. Huge amounts of clean information in Wikipedia allowed us to uncover important relations between concepts and suggest excessive low volume, inexpensive keywords. Also, we will show the viability of our approach by comparing its results to recent proposed systems. Compared to previous researches, our proposed approach have many advantages, namely, being language independent, being well-grounded, containing expert keywords and being more computationally efficient.

Keywords: Search Engine Marketing, Sponsored Search, Keyword Generation/Suggestion, Wikipedia-Mining, Semantic Relatedness, Relevance-Based Language Models

1. Introduction

The increasing growth of the World Wide Web constantly enlarges the revenue generated by search engine advertising from 3.6 billion in 2004[1] to almost 20 billion by 2007 and 31 billion in 2011[2] with a growth rate of almost 20% year over year[3]. Sponsored search or Search Engine Marketing (SEM)[4] is a form of content advertising on the Internet where advertisers bid on keywords associated with their bushiness to display their ads alongside search results with some fee for each click on the ad[5]. The placement of these advertisements is generally related to some function of the relevance to the query and the advertiser's bid. Given a seed keyword (a single word or a short phrase), advertisers bid for the keyword, and the winners of the auction have their ads displayed as sponsored links next to the search results[6].

There is a gap between the keywords chosen by advertisers and customers. So it is important to find out new alternative keywords, relevant to the base query, but non-obvious in nature. The problem of identifying an appropriate set of keywords for a specific advertiser is called keyword suggestion/generation (or keyword research). It is an emerging new scientific sub-discipline, at the intersection of large scale text analysis, information retrieval, statistical modelling and machine learning. This problem is important for both the search engines and the advertisers. From search engine perspective the revenues from SEM exceed billions of dollars and continues to grow diligently, so all popular search engines provide services for keyword research (e.g., Google's AdWords Keyword Tool[7], Overture/Yahoo! Keyword Selector Tooland Microsoft adCenter Labs Keyword Group Detection[8]). From academic perspective, this problem poses more limitations and challenges than other related issues such as query expansion or semantic similarity between terms[9], since it combines relevance with user interaction models, advertiser valuations, and commercial constraints. For example, proposed keywords must be inexpensive, expertise and non-obvious in nature[10] and also ads should be matched to the commercial intent of the users with very much condensed information with specific linguistic features rather than to their information needs[3]. It is obvious that such a problem needs a comprehensive knowledge repository. We propose Wikipedia[11] as a multilingual, web-based, free-content encyclopaedia for this issue. One of the major strengths of Wikipedia is that it

* Corresponding author. Email: amir.jadidi@qiau.ac.ir

contains information about a specific entity in the world, not available conveniently through other resources such as WWW. So, recently, many researchers endeavour to leverage Wikipedia as a terrific knowledge repository[11, 12].

We combine state-of-the-art strategies for advertising keyword suggestion with Wikipedia's unique property. The practical motivation for our work was creating a "keyword suggestion" approach, which, for any given search query, provides a list of appropriate keyword suggestions from the Wikipedia. The contributions of this paper are as follows:

- We provide a unique approach for advertising keyword suggestion by extracting a slice of Wikipedia for a specific query based on a well-known relevance-based language model in tandem with exploiting different aspects of Wikipedia.
- We categorize different word senses for ambiguous queries by capitalizing Wikipedia's disambiguation pages. It is an important feature in our model that oversights from previous researches[9, 10,13].

The rest of the paper is organized as follows: Section 2 discusses related works. Section 3 introduces some features of Wikipedia utilized by our system and Section 4 paints a general picture of the suggestion approach. The evaluation of our keyword suggestion method will be shown in Section 5. Finally, we conclude the efforts and discuss further research possibilities in Section 6.

2. Related Works

Extracting keywords from documents is a popular problem. KEA[14] is a well-known keyword extraction tool, which employs a naive Bayes learning algorithm. Later, Medelyan etal.[15] utilized Wikipedia[11] as a controlled vocabulary for identifying key phrases in a document, with article titles serving as index terms and redirect titles as their synonyms. Query Expansion[16] is another related field. These techniques can be categorized as either global or local. While global techniques rely on analysis of a whole collection to discover word relationships, local techniques emphasize analysis of the top-ranked documents retrieved for a query[17].

The bulk of research on advertising keyword suggestion has focused on exploiting the content of either Web pages[13] or search engine results[10]. Yih et al.[13] proposed a system for extracting keywords from Web pages for contextual advertising. Among other features, they used the frequencies of candidate keywords in the query log. "Proximity-based tools" query the search engines for the seed keyword and then use the retrieved text snippets or documents to extract relevant keywords. Although this technique can generate a large number of suggestions, it cannot produce relevant keywords that do not contain the original term. Such words have a good chance of being among expensive keywords, as they are already popular in the advertising community.

To overcome this problem, recent works have raised concerns about mining semantic relationships between terms for suggestions[18, 19]. Joshi and Motwani[10] leveraged search engines to capture semantic relationships between terms as a directed graph so-called TermsNet and produces non-obvious, extendable, inexpensive keywords. Later, Abhishek and Hosanagar[9] used a web based kernel function to establish semantic similarities between terms. The similarity graph is then traversed to generate keywords that are related and cheaper. Also Jones et al.[20] generated highly relevant query substitutions with some inspiration from machine translation techniques. Categorizing different concepts for a given query is a new challenge in this area. [21]and[22] are scant researches that address keyword ambiguity in advertising.

Search engine query-click logs maintain the queries that users pose to the search engine and the documents that are clicked in return. The Google Adwords Tool[7] relies on query log and advertiser log mining for keyword suggestion. It simply presents frequent queries that contain the entire search term. These techniques suffer from drawbacks like proximity-based searches and cannot find relevant keywords not containing the exact seed query words. To overcome this problem, Fuxmn et al.[23]formulated it as a semi-supervised learning problem, and proposed algorithms within the Markov Random Field model and exploited the relationship between queries and URLs to find queries that are related to the interests of the advertiser. Also click-data have been leveraged by[24] to learning and evaluating sponsored search ranking systems.

Another method focuses on extracting new keywords from meta-tags. Many high ranked websites include relevant keywords in meta-tags. Popular online tools like WordTracker[1] use meta-tag spiders to search seed keywords and make suggestions for meta-tag words based on highly ranked web pages. Although these techniques give popular keywords closely related to the base keyword, the number of relevant keywords generated is low. So they are not good candidates for keyword research.

Studies have shown that one of the main success factors for contextual advertising is their relevance to the surrounding content. Broder et al. [25] proposed a way of matching advertisements to web pages that rely on a semantic match as a major component of the relevance score based on page classification. On the other hand, [26] proposed a framework for associating ads with web pages based on Genetic Programming.

Table 1
Attributes of Wikipedia corpus

Attribute	Value
Created date	2008-07-14
Original Corpus format	XML
Original Corpus size	17 GB
Processed Corpus size	37 GB
Processing Time	66:14:29
Number of articles	7,236,522
Number of titles	3,834,509
Number of redirects	3,038,557
Number of templates	143,827

3. Wikipedia Corpus

Although we can accumulate many gigabytes worth of textual data, but at a price, texts obtained from the Web are often plagued with noise. One of the main advantages of Wikipedia[11] is article's clarity because its articles are much cleaner than typical Web pages, and mostly qualify as standard written English. Also, it contains special features such as disambiguation pages, redirects, category hierarchy, rich link structure, etc. that distinguish it from other online resources. So it is a good candidate for the knowledge repository we are looking for. [27] found Wikipedia accuracy to rival that of Encyclopaedia Britannica.

All the articles in Wikipedia are available for download through their weekly data dumps. We keep a local copy of such data for faster access. Table 1 depicts statistics and features of Wikipedia corpus that were utilized in our experiments.

Disambiguation in Wikipedia is the process of resolving conflicts in concepts that occur when a single term can be associated with more than one topic, making that term likely to be the natural title for more than one article. For example, the word "skin" can refer to several different meanings. In the corresponding disambiguation page - http://en.wikipedia.org/wiki/Skin (disambiguation), eight different semantic groups have been discovered by authors. Capitalize Wikipedia disambiguation pages is one of the important contribution in this paper that was disregarded in the previous related researches [9,10,13].

Polysemy is one of the well-known problems in Information Retrieval and manyresearches have been conducted to solve this problem. Redirect pages are Wikipedia's solutionthat manage abbreviations/shortcuts, misspellings, other spellings, plurals, related words, etc. For example, the word "Thailand" redirects to "Siam" or "Sayam" in Wikipedia (used in "massage" query in Table 3) or the word "DNA" redirects to "Deoxyribose Nucleic Acid". Redirect pages are very rich in our snapshot of Wikipedia. More than 3 million redirects were extracted and used in our experiments.

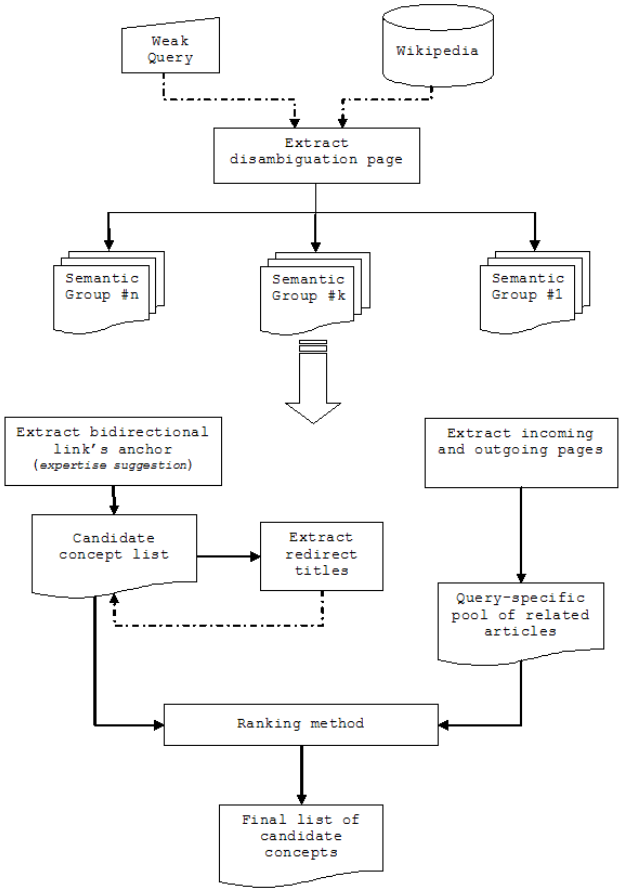

Fig.1. system architecture

It is only natural for an electronic encyclopaedia to provide cross-references in the form of hyperlinks. As a result, a typical Wikipedia article has many more links to other entries than articles in the conventional printed encyclopaedias. We leverage Wikipedia's rich links and contents for keyword suggestion problem.

4. Methodology

The task of advertising keyword suggestion in our approach can be broken into the following steps:

- Categorizing different word senses and semantic groups by leveraging Wikipedia disambiguation pages (it is only needed for ambiguous queries).
- Create an initial list of candidate concepts for a founded semantic group by tracking bidirectional link's anchors (expertise suggestion).
- Fertilizing candidate concepts by extracting redirect titles for each item in the initial candidate list. (slangy suggestions)
- Creating a pool of related articles by tracking incoming and outgoing links for a specific recognized semantic group.

- Using a well-known Relevance-Based Language Model to extract a large list of relevant suggestions.
- Ranking the candidate list of terms.

This section addresses these steps in details. As can be seen, during the experiments, two important products will be extracted: a list of candidate concepts and a pool or local-corpus (query-specific corpus) of related articles that utilizerelevance-based language model component.

4.1. Categorize Word Senses

Managing disambiguation concepts is one of the important factors that was disregarded in previous researches[9,10,13]. [21] and[22] are scant researches that address keyword ambiguity in advertising. We extract different semantic groups for a given ambiguous query by capitalizing Wikipedia's disambiguation pages. Fig 1 depicts our approach. We can handle ambiguous queries and generate different keywords in each semantic group, so the end user can select the appropriate context and then a list of suggestions will be generated by the following approach. For example, we can generate a list of suggestions for the "skin" query in various contexts such as Biology, Art and Music other than human skin. Finally, selecting the appropriate context by the end user leads our system to one (or more) kernel pages.

Disambiguation pages are distinguished in Wikipedia using "(disambiguation)" in title, so we can easily handle them. On the other hand, most Wikipedia pre-processing tools such as Wikipedia Miner [28] or JWPL[29] can handle disambiguation titles.

4.2. Extract a List of Candidate Concepts

After selecting the appropriate semantic groups (context) that are related to the user's intent, a rich list of candidate concepts will be generated by leveraging bidirectional link's anchor. As discussed in[15], links are often made to hyponyms or instances rather than synonyms of the anchor text, but unlike some applications such as Topic Indexing[15, 30], we found that hyponyms are very informative in the advertising problem. We believe that bidirectional link's anchor for a specific kernel page in Wikipedia contains expertise suggestions about it. For example, for "skin" query, the corresponding page in Wikipedia contains bidirectional links to a lot of related concepts such as: ``skin components", ``skin diseases" and etc. Note that we utilize bidirectional links, not outgoing links, because some articles are linked just for their existence[11]. Using bidirectional linksallows us to filter noisy terms. Wikipedia has a really rich link structure. Table 2 illustrates the number of various links for a given weak query in our experiments.

Although currently suggestions for each weak query are almost enough for some applications, one of the main challenges in advertising keyword suggestion is that hundreds or sometimes thousands of relevant keywords must be generated.

Table 2

Extracted features from Wikipedia articles for each benchmark query

Concept Features	Skin	Teeth	Pedicure	Massage
incoming links	1993	1254	20	596
outgoing links	314	521	20	564
redirects	2212	2655	113	3969
candidate concepts	2526	3176	133	4533
pool size	2307	1775	40	1160

To overcome this problem and generate more suggestions, we offer Wikipedia's redirect pages. Our experiment showed that by leveraging redirect titles we can generate a huge list of related concepts. For example, the initial candidate list for "skin" query contains 314 suggestions that by leveraging redirect titles for each of them, 2212 slang or equivalent concepts will be generated.

4.3. Extract Query-Specific Pool of Related Articles

Although the mechanism described in Section 4.2 can generate a huge list of suggestions, we need an approach to categorize them by importance and relevance. To overcome this problem, we offer a well-known relevance-based language model to generate a ranked list of general seed keywords and then categorize our candidate concepts by them. Lavrenko and Croft[31] proposed an approach for estimating a relevance model with no training data. They used the term ``relevance model" to refer to a mechanism that determines the probability $P(w|R)$ of observing a word w in the documents relevant to a particular information need. This method was applied to different related areas such as query expansion successfully[31], [32]. As relevant-based LM method is a statistical method for query expansion hence the proposed keywords will begeneral.

Table 3 compares different systems for a series of queries evaluated by previous works[9], [33]. Although Lavrenko's relevance model does not apply to the area of advertising, the results of it are comparable to previous works. We believe this is because of Wikipedia article's clarity that make an occasion for a statistical method.

5. Experiments

First of all, Wikipedia original XML dump was processed to extract different necessary features. Table 1 depicts the statistics of our version of Wikipedia. Lavrenko's relevance model [31]was used with 10 top documents and maximum 5-grams for proposed term's length. Same parameters were used for whole corpus and query-specific variant. We leveraged 2008-07-14 offline XML version of Wikipedia. Using Wikipedia-Miner [28], an open-source tool for mining Wikipedia, more than 7 million articles

wereextracted. These articles are used to build the relevance-based language model.

Table 3

Comparing Lavrenko's relevance model[31] on the whole Wikipedia collection addressed in Table 1 and query-specific relevance-based language model described in Section 4.3 with Wordy[9] and Wikipedia Concept Graph (CG)[33].

Case Study: Comparison				
Query	Wordy	Concept Graph	**Query-Specific LM**	**Corpus LM**
Skin	skincare	psoriasis	skin	skin
	facial	Inhale	human	care
	treatment	epidermis	hair	america
	face	uvb	body	use
	care	danger	cell	skin care
	occitane	corneum	light	organ
	product	melanocytic	lay	natural
	exfoliator	harm	africa	human
	dermal	exposure	dark	product
	body	prolong	people	form
Teeth	tooth	tooth	teeth	teeth
	whitening	xtract	deciduous teeth	tooth
	dentist	dentition	permanent teeth	shark
	veneer	dentist	molar	molar
	filling	orthodontic	age	permanent
	gums	enamel	develop	horse
	face	incisor	tooth	mouth
	baby	dental	incisor	develop
	smilesb.	premolar	mouth	incisor
	features	molar	dental	dental
Pedicure	manicure	-	pedicure	pedicure
	leg	-	nail	nail
	feet	-	progress	foot
	nails	-	use	use
	treatment	-	foot	feet
	skincare	-	feet	massage
	tool	-	time	progress
	smilesb.	-	treat	stone
	massage	-	massage	manicure
	facial	-	salon	care
Massage	therapy	heritage	massage	massage
	therapy	therapist	thailand	therapy
	massagea.	knead	associate	therapist
	therapist	parlor	massage th.	thailand
	therapeutic	kahuna	massage th.	use
	thai	erotic	thailand msg.	associate
	oil	reflexology	bodywork	massage th.
	bath	perineal	profession	chair
	offer	therapy	tradition	massage th.
	styles	shiatsu	service	prostate

In our experiments, we consider queries used in[9] and compare the results of our system with the results of Wordy[9] and Concept Graph[33] on the same queries. Table 3 compare our proposed model based on Lavrenko's relevance model[31] on the whole Wikipedia corpus and query-specific corpus contains related articles described in Section 4.3 with Wordy[9] and Concept Graph[33]. We extracted a list of candidate concept as described in Section 4.2 and used Lavrenko's relevance model results to categorize our candidate concept list.

Although in Table 3, only top ranked terms are shown, analysis of the results reveals that our method suggests more professional keywords. For example very technical terms such as ``epidermis", ``uv" and ``melanin" are available at the lower position in the ranked list. It means that the general keywords get more weight in the Lavrenko's relevance model[31]. To show the commonness of the proposed keywords in Table3, some lower rank items of our results for "skin" query are shown in Table4. On the other hand, our proposed method extracts both general and specific keywords for a given query. Since the underlying rank function is a raw statistical function, general keywords are shown in the top of the ranked list, while more specific keywords are in the bottom. However, there are several ranking methods proposed in the literature that utilize common IR measures such as term frequency but Wikipedia has special features such as depth of the category hierarchy that can be utilized to balance commonness and relatedness factors in the final ranking method. Using these two factors, the proposed system can be tuned with the users' offers. It is very important in the domain of advertising because general terms are usually expensive but pervasive, while specific terms are inexpensive and professional. Note that both general and specific keywords are important in Search Engine Marketing.

Weak and ambiguous queries are well-known between users of Web Search Engines. Unfortunately, Wordy case studies (test cases in Table3) are so clear and unambiguous. As pointed out in Section 1, detecting different contexts of the given query is a distinguishing characteristic of our proposed method that is completely neglected in Wordy[9] and Concept Graph[33]. Using a more comprehensive benchmark dataset that contains ambiguous queries is a better evaluation metric in the real world.

Previous researches in SEM show that the incoming queries to a search engine are completely related to the newsworthy events. Google Trends is a well-known tool in this regard. Wikipedia is a collaborative knowledge repository that is completely sensible to the world events. It means that Wikipedia-based suggestion systems can be sensible to the newsworthy events as well.

6. Conclusion and Future Work

We have introduced an approach to Advertising Keyword Suggestion that leverage the knowledge available in the Wikipedia. Among different researches [9, 10, 12, 13, 33], our proposed model is one of the scant researches has been utilized different aspects of Wikipedia such as disambiguation pages, redirects pages, cross-reference and link structures. Since the perfect evaluation of the proposed model needs bids data, we conducted a series of case studies to evaluate and compare impact of our approach with other related researches[18].

There are several directions for future work. First, we must use an appropriate ranking function. Also

incorporating keyword pricing information in ranking suggestions can be useful to produce a lower price term suggestion [2, 5,6].

Table 4

Low rank keywords extracted by our proposed system.

Low rank keywords for a sample query: "skin"		
Query	**Corpus LM**	**More Suggestions**
Skin	epidermis	epidermis, lower epidermis, upper epidermis
	gene	genes, epigenetically, gene structure, epigenetic, cancer genet. cytogenet., epigenetic regulation, epigene, epigenetic inheritance, gene sequence, epigenetic imprinting, melanogenesis, ...
	vitamin	vitamin d, vitamin d deficiency, b vitamin, vitamin b100, b complex vitamins, b-vitamins, vitamin d-binding protein, vitamin d excess, vitamin b complex, vitamin-d, ...
	uv	uv, uva radiation, extreme uv, uv a, uv-a, vacuum uv, uv rays, uv-radiation, uvb radiation, uv-b, uv ray, ...
	melanin	melanin, pheomelanin, brown eumelanin, neuromelanin, melanins, eumelanin, phaeomelanin, ...
	sun	sunsclad, sunlit, sun bathing, sun scald, sun exposure, sunbath, sunscald, mvemjsun, ...
	dermatitis	dermatology, list of dermatological diseases, transdermal drug delivery system, epidermal, dermal ridge, ...

Second, Wikipedia category hierarchy is an important piece of knowledge available to us. More suggestions will be produced by traversing category hierarchy for specific concepts. Intensive researches have been conducted to utilize Wikipedia category hierarchy in different areas of knowledge discovery and information retrieval.

Third, one of the important features in Wikipedia is the multilingual content. At present, we have some Chinese keywords in our results that come from redirects and anchors. By leveraging multilingual contents of Wikipedia we can generate cross-language advertising that can be a valuable contribution to the literature.

Last, one important direction is to consider more strict evaluation frameworks. Although we conducted our experiments based on previous case studies, we need a benchmark dataset to evaluate and compare our systems with the related researches.

References

[1] Cristo, M., et al., Search Advertising, in Soft Computing in Web Information Retrieval, E. Herrera-Viedma, G. Pasi, and F. Crestani, Editors. 2006, Springer Berlin Heidelberg. pp. 259-285.

[2] Thomaidou, S., K. Leymonis, and M. Vazirgiannis, GrammAds: Keyword and Ad Creative Generator for Online Advertising Campaigns, in Digital Enterprise Design and Management 2013, P.-J. Benghozi, D. Krob, and F. Rowe, Editors. 2013, Springer Berlin Heidelberg. pp. 33-44.

[3] Dominowska, E. and V. Josifovski. First Workshop on Targeting and Ranking for Online Advertising. in Proceedings of the 17th International Conference on World Wide Web. 2008. New York, NY, USA: ACM.

[4] Ghose, A. and S. Yang, An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets. Management Science, 2009. 55(10): pp. 1605-1622.

[5] Blankenbaker, J. and S. Mishra, Paid search for online travel agencies: Exploring strategies for search keywords. J Revenue Pricing Manag, 0000. 8(2-3): pp. 155-165.

[6] Blask, T., B. Funk, and R. Schulte, To Bid or Not To Bid? Investigating Retail-Brand Keyword Performance in Sponsored Search Advertising, in E-Business and Telecommunications, M. Obaidat, J. Sevillano, and J. Filipe, Editors. 2012, Springer Berlin Heidelberg. pp. 129-140.

[7] Geddes, B., Advanced Google AdWords. 1st ed. 2010, Alameda, CA, USA: SYBEX Inc.

[8] Seda, C., Search Engine Advertising: Buying Your Way to the Top to Increase Sales. 2004, Thousand Oaks, CA, USA: New Riders Publishing.

[9] Abhishek, V. and K. Hosanagar. Keyword Generation for Search Engine Advertising Using Semantic Similarity Between Terms. in Proceedings of the Ninth International Conference on Electronic Commerce. 2007. New York, NY, USA: ACM.

[10] Joshi, A. and R. Motwani. Keyword Generation for Search Engine Advertising. in Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on. 2006.

[11] Medelyan, O., et al., Mining meaning from Wikipedia. Int. J. Hum.-Comput. Stud., 2009. 67(9): pp. 716-754.

[12] Zhang, W., et al., Advertising Keywords Recommendation for Short-Text Web Pages Using Wikipedia. ACM Trans. Intell. Syst. Technol., 2012. 3(2): pp. 1-25.

[13] Yih, W.-t., J. Goodman, and V.R. Carvalho. Finding Advertising Keywords on Web Pages. in Proceedings of the 15th International Conference on World Wide Web. 2006. New York, NY, USA: ACM.

[14] Frank, E., et al. Domain-Specific Keyphrase Extraction. in Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. 1999. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[15] Medelyan, O., I.H. Witten, and D. Milne. Topic indexing with Wikipedia. in Proceedings of the Wikipedia and AI workshop at AAAI-08. 2008. AAAI.

[16] Milne, D.N., I.H. Witten, and D.M. Nichols. A knowledge-based search engine powered by wikipedia. in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007. New York, NY, USA: ACM.

[17] JadidiNezhad, A. and H. Amiri, Local cluster analysis as a basis for high-precision information retrieval. The 6th International Conference on Informatics and Systems, INFOS, 2008.

[18] Gabrilovich, E. Ad retrieval systems in vitro and in vivo: knowledge-based approaches to computational advertising. in Proceedings of the 33rd European conference on Advances in information retrieval. 2011. Berlin, Heidelberg: Springer-Verlag.

[19] Kim, C., et al., An empirical study of the structure of relevant keywords in a search engine using the minimum spanning tree. Expert Systems with Applications, 2012. 39(4): pp. 4432-4443.

[20] Jones, R., et al. Generating Query Substitutions. in Proceedings of the 15th International Conference on World Wide Web. 2006. New York, NY, USA: ACM.

[21] Wu, X. and A. Bolivar. Keyword Extraction for Contextual Advertisement. in Proceedings of the 17th International Conference on World Wide Web. 2008. New York, NY, USA: ACM.

[22] Chen, Y., G.-R. Xue, and Y. Yu. Advertising Keyword Suggestion Based on Concept Hierarchy. in Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008. New York, NY, USA: ACM.

[23] Fuxman, A., et al. Using the Wisdom of the Crowds for Keyword Generation. in Proceedings of the 17th International Conference on World Wide Web. 2008. New York, NY, USA: ACM.

[24] Ciaramita, M., V. Murdock, and V. Plachouras. Online Learning from Click Data for Sponsored Search. in Proceedings of the 17th International Conference on World Wide Web. 2008. New York, NY, USA: ACM.

[25] Broder, A., et al. A Semantic Approach to Contextual Advertising. in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2007. New York, NY, USA: ACM.

[26] Lacerda, A.i., sio, et al. Learning to Advertise. in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006. New York, NY, USA: ACM.

[27] Giles, J., Internet encyclopaedias go head to head. Nature, 2005. 438(7070): p. 900-901.

[28] Milne, D. and I.H. Witten, An open-source toolkit for mining Wikipedia. Artificial Intelligence, 2013. 194(0): pp. 222 - 239.

[29] B\"a, r., Daniel, T. Zesch, and I. Gurevych. DKPro Similarity: An Open Source Framework for Text Similarity. in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2013. Sofia, Bulgaria: Association for Computational Linguistics.

[30] Kim, S., et al., Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation, 2013. 47(3): pp. 723-742.

[31] Lavrenko, V. and W.B. Croft. Relevance Based Language Models. in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001. New York, NY, USA: ACM.

[32] Diaz, F. and D. 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. 2006. New York, NY, USA: ACM.

[33] H. Amiri, A.A., M. Rahgozar and F. Oroumchian, Keyword Suggestion Using Concept Graph Construction from Wikipedia Rich Documents, 2008.