Quality Evaluation of Search Results by Typicality and Speciality of Terms Extracted from Wikipedia

Makoto Nakatani, Adam Jatowt, Hiroaki Ohshima, and Katsumi Tanaka

Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-8501 Japan
{nakatani,adam,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. In Web search, it is often difficult for users to judge which page they should choose among search results and which page provides high quality and credible content. For example, some results may describe query topics from narrow or inclined viewpoints or they may contain only shallow information. While there are many factors influencing quality perception of search results, we propose two important aspects that determine their usefulness, "topic coverage" and "topic detailedness". "Topic coverage" means the extent to which a page covers typical topics related to query terms. On the other hand, "topic detailedness" measures how many special topics are discussed in a Web page. We propose a method to discover typical topic terms and special topics terms for a search query by using the information gained from the structural features of Wikipedia, the free encyclopedia. Moreover, we propose an application to calculate topic coverage and topic detailedness of Web search results by using terms extracted from Wikipedia.

Keywords: Search results quality, Wikipedia mining, Term extraction, Term typicality, Term speciality.

1 Introduction

Web search engines have become frequently used for acquiring information over the Internet. Web search results given by search engines are usually composed of a list of Web pages with some information such as titles, snippets and urls. However, it is often difficult for users to judge which page they should choose among search results. In many cases, users require Web pages including credible and comprehensive information about the search query. According to the online survey that we have recently conducted on 1000 respondents in Japan [1], users search the Web mostly because they require basic (46%) or detailed (36.8%) information about their search queries. Yet, conventional search engines usually do not provide users with any detailed information about the extent to which search results cover typical query topics. Some Web pages in search results may be regarded as being of low quality because they contain information related to query topics that are described from a narrow or an inclined viewpoint. In this

X. Zhou et al. (Eds.): DASFAA 2009, LNCS 5463, pp. 570–584, 2009.

sense, a page is deemed to be of high quality if it covers as many typical topics about a query term as possible. On the other hand, if a Web page conveys only shallow information in spite of covering many typical topics, the page will be also regarded as low quality one. In this paper, we propose the notion of the "topic coverage" and "topic detailedness" of Web pages for evaluating their quality. Topic coverage of a Web page means how many typical topics about the search query are covered by the Web page. On the other hand, topic detailedness of a Web page intuitively means how many special topics are included in the page. We believe that it will become easier for users to judge which page they should choose by showing them the above two measurements.

We would like to emphasize here that the complete quality evaluation of web pages is actually a complex, multi-dimensional issue. In order to find high quality pages one would generally have to analyze many aspects such as information accuracy and freshness, content organization, completeness, readability and so on. In this research we focus only on two aspects of quality evaluation of Web pages, topic coverage and topic detailedness. Both are actually query-dependent quality measures and can thus fit well into a search scenario in which users seek high quality pages for their queries. The proposed measures are also user-dependent to some extent. For example, users who are experts within certain topics would probably search for highly-specialized, detailed pages while non-experts users should generally prefer documents covering broad and typical topics related to their queries.

For calculating the topic coverage and topic detailedness of Web search results it is first necessary to extract typical and special terms for a search query used for generating these results. In this paper, we define typical and special terms for a search query as follows:

- Typical terms of a search query are terms that frequently appear in the domain of the search query.
- Special terms of a search query are terms that appear mostly in the domain of the search query.

Typical terms are used for measuring topic coverage of Web pages, and special terms are used for measuring topic detailedness.

We propose a method to discover typical topic terms and special topics terms for a search query by using the information gained from the structural features of Wikipedia. Wikipedia, the free online encyclopedia that anyone can edit, provides a huge number of interlinked entries. It started in 2001 becoming a prominent example of successful collaboration of thousands users on the Web. According to the statistics which Wikipedia has released as of June 2008 [1], the English Wikipedia contains about 2.4 million articles, and there are about 7 million registered user accounts. According to the Nature Journal, Wikipedia is about as accurate in covering scientific topics as the Encyclopedia Britannica [2]. In this work, we focus on category and link structure of Wikipedia for the purpose of measuring typicality and speciality of terms. In Wikipedia, each

[1] http://en.wikipedia.org/wiki/Special:Statistics

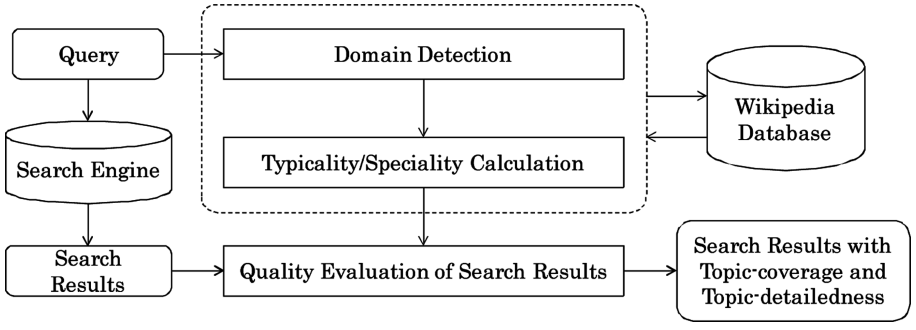

Fig. 1. Overview of our proposed system

article is assigned to one or more categories, and it links to and is linked by other related articles. In our approach, the category structure is used for detecting the domain of query term, and we calculate typicality and speciality of topic terms by analyzing the link structure in Wikipedia.

We have also implemented a system that presents Web search results with the scores of topic coverage and topic detailedness. The overview of the proposed system is illustrated in Fig. 1. Given a query, (i) the domain of a search query is detected, and (ii) typicality and speciality scores of terms are calculated by using the category and link structure of Wikipedia. At the last, (iii) topic coverage and topic detailedness of each Web page acquired from a search engine are measured by using typical and special terms extracted from Wikipedia and pages are annotated with these both measures.

The remainder of the paper is organized as follows: Section 2 discusses related work. In Section 3, we propose the method of measuring typicality and speciality of terms by using the structural features of Wikipedia. In Section 4, we present the approach of measuring topic coverage and topic detailedness of Web search results by using typicality and speciality of terms. Section 5 provides conclusion and discusses our future work.

2 Related Work

2.1 Quality Evaluation of Web Pages

The quality of Web pages has been evaluated so far from various viewpoints. Link analysis has been probably the most frequently exploited approach for the quality evaluation in information retrieval. PageRank [3] and HITS [4] are well-known algorithms in which the number of in-links of a Web page are used as a rough measure for the popularity and, indirectly, the quality of the page. Following the success of PageRank, Haveliwala [5] proposed a topic-sensitive PageRank measure, which separately determines a set of popularity scores for predetermined topics. Cho et al. [6] discovered that page ranking by link analysis causes the "rich-get-richer" phenomenon, and they proposed the method of

measuring quality from Web snapshots by analyzing the changes in PageRank values over time. While link analysis considers the perspective of Web page authors, the information extracted from social annotations generated by users has recently attracted much attention for evaluating Web contents. The possibility of evaluating the quality of Web pages by using the information extracted from social bookmarking sites such as Del.icio.us[2] is described in [7][8].

Some researchers also proposed machine learning approaches for evaluating the quality of Web pages [9][10][11]. In these approaches, HTML structure, the number of links and language features such as number of unique words and so on are used as parameters for machine learning. Mandl et al. [11] implemented AQUAINT, a quality-based search engine, using a machine learning method. Our method is different from these works in that it uses Wikipedia, as a knowledge base constructed by the collaborative effort of multiple users. We also propose two query-dependent factors for page quality measurement, topic coverage and topic detailedness by which our method analyzes Web pages.

2.2 Term Extraction

Large text corpora have been successfully used for knowledge extraction. For example, Hearst [12] proposed a method for the automatic acquisition of the hyponymy lexical relations from unrestricted text. Several researchers have begun to seek effective ways for mining huge data collections since the detailed analysis of large content repositories is often impossible or prohibitively costly. Bollegala [13] proposed a semantic similarity measure that uses page counts and text snippets returned by a Web search engine for computing the similarity between terms or entities. In a similar fashion, Cilibrasi and Vitanyi [14] introduced a semantic distance measure called Google Normalized Distance between query terms based on the returned Web count values.

Wikipedia has recently attracted much attention as a large-scale, semi-structured corpus for data mining; and "Wikipedia mining" has become a new research area [15][16][17]. Strube [15] proposed a method of measuring semantic relatedness by using category data of Wikipedia articles. In addition, several applications based on knowledge extracted from Wikipedia have been demonstrated [18][19]. For example, *Koru* [18] is a new search engine that uses knowledge from Wikipedia for automatic query expansion. The Wikify! system proposed by Mihalcea [19] attaches links to entry articles of terms selected in Web pages by using the keyword extraction and word sense disambiguation based on Wikipedia data. These systems help users with search and learning, while our system aims at evaluating quality of Web pages by using the information gained from the Wikipedia.

Our work is also related to detecting domain-specific knowledge. Some methods for extracting domain-specific terms from online documents have been proposed in various domains, such as, the field of medical informatics [20][21]. Eibe et al. [22] described a method for finding domain-specific phrases by using

[2] http://del.icio.us

machine learning techniques. The above solutions, however, usually require a large number of manually labeled training data. Bing et al. [23] introduced a non-supervised method for detecting topic-specific concepts and definitions from Web pages. Their techniques first identify sub-topics or salient concepts of the topic, and then find and organize informative pages to be presented for users. Our approach differs from these methods in that, first, it extracts domain-specific terms from Wikipedia for the purpose of quality evaluation and, second, it is based on unsupervised and domain-independent algorithm.

3 Typicality and Speciality of Terms

There are many studies about term extraction as mentioned in the above section. Yet, typicality or speciality of extracted terms have not been referred in those works. They are prerequisite for assessing coverage and detailedness of Web pages returned for a query.

Our proposed method is composed of the following steps. To extract typicality and speciality of topic terms we first detect the domain of a search query by using the link and category structure of Wikipedia. The next step is to extract terms from Wikipedia articles included in the detected domain. After term extraction, we calculate typicality and speciality of terms by analyzing the distribution of links to the article of each term.

3.1 Detecting a Domain of Search Query

First, we describe the method for detecting a domain of a search query when there exists a Wikipedia article about the original query term. Suppose that q is a search query and a_q is a Wikipedia article about q. The first step is to acquire the set of categories that a_q belongs to. We express it as $C_{direct}(q)$. Each category that $C_{direct}(q)$ contains is intuitively a hypernym of the query term. For example, an article about "iPod" belongs to ten Wikipedia categories such as "Portable Media Player", "2001 introductions" and "Semi-protected" etc. "Portable Media Player" is regarded as an appropriate category for C_{direct}("iPod"). Although "2001 introductions" can be regarded as a hypernym of "iPod", we remove it from direct categories since other articles contained in this category are hardly related to "iPod". "Semi-protected" is also removed from direct categories because it is a label that expresses the temporal state of articles. In the proposed method, we do not deal with categories that are classified by time axis such as "20xx_yyyy" and with the categories which are only labels about article state such as "Protected" and "All articles with unsourced statements" etc.

We consider that not only direct categories but also indirect categories are required for expressing in what kind of context a search query is used. In the example of "iPod", categories such as "iPod software" and "iPod accessories" are not directly combined with the article about "iPod". However, these categories are strongly related to "iPod". In order to add such indirect categories to the domain of a search query, we gather the category data of articles that are linking

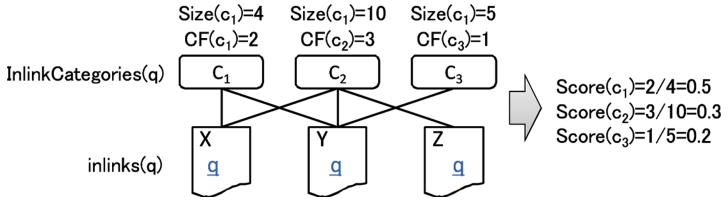

Fig. 2. Explanation of indirect category scoring where inlinks(q) is the articles linking to the article of q

to an article about the query. A basic idea about how to find indirect categories is described in Fig. 2. First, articles linking to an article about a search query, expressed as $inlinks(q)$, and categories that those articles belong to (denoted as $InlinkCategories(q)$) are acquired:

$$InlinkCategories(q) = \cup_{a_i \in inlinks(q)} Categories(a_i) \tag{1}$$

Next, the categories containing many articles from $inlinks(q)$ could be regarded as indirect categories. However we need to consider the size of each category. The score of each category in $InlinkCategories(q)$ is measured by the following equation:

$$Score(c) = \frac{CF(c)}{Size(c)} \tag{2}$$

where $CF(c)$ is the number of articles which are contained both in a category c and in $inlinks(q)$, and $Size(c)$ means the number of articles contained in the category c. If $Score(c)$ is larger than a threshold α and the category contains more than β articles, the category c can be regarded as belonging to $C_{indirect}(q)$, indirect categories of query. In this paper, α is set as 0.5 and β is set as 5. Next, $Domain(q)$, a domain of search query, is determined by calculating the union of direct categories and indirect categories:

$$Domain(q) = C_{direct}(q) \cup C_{indirect}(q) \tag{3}$$

In case when a Wikipedia article about the original query does not exist, the query is divided into as long word sequences ($q = \{q_1, q_2, ..., q_n\}$) as possible for which Wikipedia articles exist. For example, if "iPod nano battery" is given as a query, it can be divided into "iPod nano" and "battery". Although the articles of "iPod" and "nano" exist, "iPod nano" is a longer word sequence than each of them and there is also a Wikipedia article about "iPod nano". Note that our proposed method cannot deal with a query for which none of the terms exist on Wikipedia articles. However, such situation rarely happens according to our study. For each of divided terms, direct categories can be extracted and category scores can be calculated in the same way as the domain detection of a single term that was described above. Direct categories for each divided terms are just adopted to as total direct categories. In order to acquire total indirect

Table 1. Examples of detecting a domain of a search query

Query:q	$C_{direct}(q)$	$C_{indirect}(q)$
iPod	iPod	iPod games
	Portable media players	Macintosh all-in-ones
		Directors of Apple Inc.
		Macintosh computers by product line
		iMac series
		iPod accessories
		iPod software
		X86 Macintosh computers
parkinson's disease	Aging associated diseases	Pervasive developmental disorders
	Geriatrics	Motor neuron disease
	Parkinson's disease	Cognitive disorders
		Tardive dyskinesia
		Dopamine agonists
		Antiparkinsonian agents
		Heterocyclic compounds (4 or more rings)
iPod headphone	Headgear	iPod accessories
	Headphones	iMac series
	iPod	
	Portable media players	

categories, category scores for each divided term are linearly combined:

$$TotalScore(c) = \sum_{q_i \in q} weight(q_i) \cdot Score_{q_i}(c) \tag{4}$$

where $weight(q_i)$ is calculated as follows:

$$weight(q_i) = \frac{w_{q_i}}{\sum_{q_i \in q} w_{q_i}} \tag{5}$$

$$w_{q_i} = \frac{\log\left(1 + |outlinks(q_i)|\right)}{\log\left(1 + |inlinks(q_i)|\right)} \tag{6}$$

Wikipedia articles for which the number of out-links is low and the number of in-links is high tend to be abstract words with broader concepts. The above weighting scheme assigns a relatively small weight value for such words. For example, if a given query is "iPhone Japan", categories of "Japan" are not important and $weight$("Japan") has a low value. To the contrary, given "iPod Zune" as a query, both terms are important and almost equivalent weights are given. The rest is the same as that in the case of a single word.

Some examples of detecting a domain of a search query are described in Table 1.

3.2 Calculating Typicality and Speciality of Terms

We describe here the way for calculating typicality and speciality scores of terms by using the link structure of Wikipedia. Intuitively, typical terms should frequently occur in the domain of a search query, while special terms should occur mostly in the domain and rarely outside of it. We explain our idea in Fig. 3. Given q as a search query, the domain of q is detected by the method described in Section 3.1. We regard terms linked by many articles included in the domain

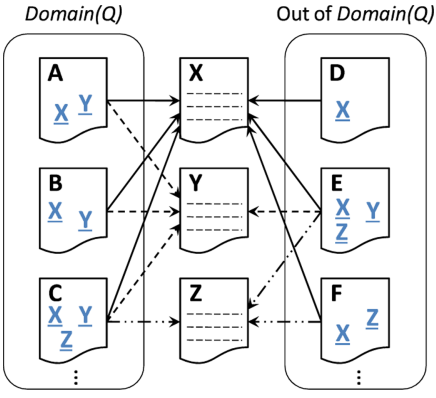

Fig. 3. Analysis of the distribution of link frequency for measuring typicality and speciality of terms

of query as typical terms, and terms linked by mostly articles included in the domain as special terms. Intuitively, a typical term is frequently used in the domain of a query, and a special term is hardly used out of the domain of query. In Fig. 3, X and Y are typical terms and Z is not a typical term. On the other hand, only Y is a special term, while X and Z are not special terms because the articles of X and Z are linked by many articles not included in the domain of the query.

Details of our proposed method are described as follows. For each category included in $Domain(q)$, we acquire all articles in it and express these articles as D_q which means domain pages of a search query q. Next, we define link frequency (LF). $LF(t, D)$ is the number of articles which link to the article of t and are included in the article set D. Typicality and speciality of a given term t when q is a search query are calculated by using $LF(t, D_q)$. Typicality score is calculated by dividing $LF(t, D_q)$ by the number of articles included in D_q, and speciality score is $LF(t, D_q)$ divided by $LF(t, D_W)$ where D_W means all articles of Wikipedia. Each equation is shown as follows:

$$Typicality(t, q) = \frac{LF(t, D_q)}{|D_q|} \tag{7}$$

$$Speciality(t, q) = \frac{LF(t, D_q)}{LF(t, D_W)} \tag{8}$$

3.3 Experiments

Experimental Setting. We prepared 20 queries in total for evaluating our proposed method of measuring typicality and speciality of terms. 10 of these queries are common terms chosen from the most popular Wikipedia articles [3].

[3] http://en.wikipedia.org/w/index.php?title=Wikipedia:Popular_pages

Table 2. Examples of typicality and speciality of terms extracted from Wikipedia

Query:q	Term:t	$Typicality(q,t)$	$Speciality(q,t)$
iPod	Apple Inc.	0.4046	0.0367
	IPod shuffle	0.2948	0.3953
	IPod Camera Connector	0.1445	0.4717
Global warming	Carbon dioxide	0.2702	0.0848
	Greenhouse gas	0.2561	0.2740
	Bali roadmap	0.1293	0.8279
Support vector machine	Machine learning	0.3755	0.3032
	Algorithm	0.1119	0.0203
	Kernel trick	0.0361	0.4545
Query expansion	Information retrieval	0.3947	0.2616
	Recall	0.0526	0.6667
	Relevance feedback	0.0439	0.833

The other 10 queries are technical terms about data mining and information retrieval, and so on. In this experiment, we do not deal with queries for which a Wikipedia article does not exist such as "iPod Zune" although in Section 3.1 we described how to process such queries. Wikipedia can be downloaded [4], and in our experiment we used the English Wikipedia database dumped in July 2008. For each query, we calculated typicality and speciality of terms by using structural features of Wikipedia. We took top 10 typical terms together with another 10 terms selected at random for each query and we showed these terms to 5 evaluators who are graduate students in informatics. Evaluators rated each term on a scale of 1 (low) to 7 (high) to reflect its typicality and speciality levels. The purpose of this evaluation is investigating the following two points:

- Accuracy of top 10 typical terms.
- Validity of speciality of each term.

Results. First, we describe some examples of typicality and speciality of terms extracted from Wikipedia in Table 2. In the example of "global warming", "carbon dioxide" and "greenhouse gas", which are generally considered as the cause of global warming, had high typicality. "Carbon dioxide" which is a relatively general term had low speciality, but "greenhouse gas" which conceptually contains "carbon dioxide" showed high speciality. This is because "greenhouse gas" is a special term only used in the domain of global warming. "Bali roadmap", a roadmap adopted after a climate change conference held in Bali, showed much higher speciality.

Next, we describe the accuracy of top 10 typical terms for our queries. We calculated an average precision for evaluating the accuracy of typicality. Here we regarded terms of which typicality score provided by evaluators were more than 5.0 as the correct terms and those with the score lower than 5.0 as incorrect ones. The result is shown in Table 3. If a technical term was given as a search query, the average precision was about 70%. On the other hand, the average precision for common term queries was only about 50%.

[4] http://download.wikimedia.org

Table 3. Accuracy of the top 10 typical terms (left table) and validity of speciality of each term (right table)

Query Type	Avg. Precision (top 10)
Common Terms	0.49
Technical Terms	0.71
Total	0.60

Query Type	Spearman's coefficient
Common Terms	0.40
Technical Terms	0.45
Total	0.42

Last, we describe the validity of the speciality calculation of the terms. We calculated Spearman's rank correlation coefficient between speciality scores measured by our proposed method and average scores given by user evaluation. As shown in Table 3, it turned out as a result that they have a positive correlation.

4 Quality Evaluation of Search Results

In this section, we propose two measurements, "topic coverage" and "topic detailedness", for facilitating the evaluation of the quality of Web pages by using typicality and speciality scores of included terms.

4.1 Topic Coverage

Topic coverage of a Web page means how many typical topics about a search query are covered by the Web page. We consider that a Web page containing more typical terms has higher coverage. Topic coverage of a Web page p about a query q is calculated by the following equation:

$$TopicCoverage(p,q) = \sum_{t \in terms(q)} C(t,p) \cdot Typicality(t,q) \tag{9}$$

where $terms(q)$ are extracted terms and $C(t,p)$ is an indicator taking values 0 or 1 depending whether a Web page p contains a given term t. A Web page with high topic coverage can be regarded as of high quality while a Web page with low coverage may be written from a narrow viewpoint or an inclined viewpoint. Note that we use here $C(t,p)$ instead of the term frequency of t in a Web page for measuring topic coverage because term frequency is not directly related to how many topics the page includes.

4.2 Topic Detailedness

Topic detailedness of a Web page means how many special topics about a search query are included in the Web page. We consider that a Web page containing more special terms is more detailed. Topic detailedness of a Web page p about a query q is calculated by the following equation:

$$TopicDetailedness(p,q) = \sum_{t \in terms(q)} TF(t,p) \cdot Speciality(t,q) \tag{10}$$

where $terms(q)$ are extracted terms and $TF(t, p)$ means term frequency of t in the Web page p. In this case, term frequency of t should be taken into consideration because Web pages with special terms that are repeatedly used are more likely to be about detailed topics.

Even if a Web page shows a high coverage, it may contain only shallow information. In general, we consider a Web page with low coverage and low detailedness as a low quality Web page.

4.3 Application

We implemented a system that presents Web search results with the scores of topic coverage and topic detailedness. The objective of this system is to make it easier for users to choose the right pages when exploring search results. Fig. 4 shows a screen shot of our system. The interface for inputting queries is the same as the one in a standard search engine. The system gives users search results which contain the scores of topic coverage and topic detailedness of each Web page in addition to titles, snippets and urls. We used Yahoo! Web search API service [5] for acquiring the search results.

Currently, our system downloads each Web page for calculating topic coverage and topic detailedness. Response speed of the system could be improved by using only title and summary for calculating the two measurements. However, we need to be aware of that it is difficult to evaluate the quality of a Web page by using only titles and snippets contained in Web search results as the length of available text is limited.

4.4 Experiments

For showing the effectiveness of our proposal, we have to clarify the following two points:

- Accuracy of our proposed method for measuring topic coverage and topic detailedness.
- Relation between the overall quality of Web pages, and topic coverage or topic detailedness.

We prepared 10 search queries for clarifying the above points. For each query, we imposed the following tasks on evaluators:

1. Read the top 10 Web pages acquired by a Web search engine.
2. Rate on a scale of 1 to 10 the topic coverage and topic detailedness of each page.
3. Rate on a scale of 1 to 10 the overall quality of each page.

In this experiment, we regarded the average of each score given by 5 evaluators as the answers of coverage, detailedness and overall quality for each page.

[5] http://developer.yahoo.com/search/web/V1/webSearch.html

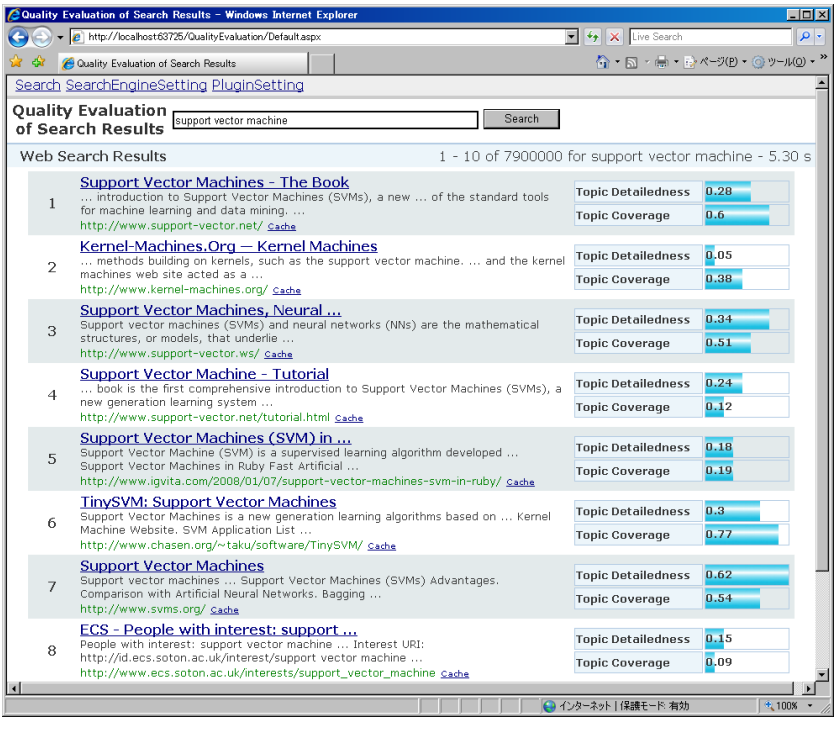

Fig. 4. A screenshot of the proposed system which presents Web search result with topic coverage and topic detailedness

We first discuss the effectiveness of our proposed method for measuring topic coverage and topic detailedness of Web pages. We used the following three rankings:

Original Ranking (OR) : An original ranking by Web search engine.
System Ranking (SR) : A ranking sorted by topic coverage or topic detailedness which we proposed.
User Ranking (UR) : A ranking sorted by topic coverage or topic detailedness scored by evaluators.

For each query, we calculated Spearman's rank correlation coefficient between the original ranking and user ranking, and between the system ranking and user ranking. The result is shown in Table 4. In general, the system ranking has more strongly positive correlation with the user ranking in comparison with the original ranking. This indicates that the proposed method for measuring topic coverage and topic detailedness is appropriate and these two measurements should help users with judging which pages they choose among search results.

Next, we investigate the correlation between topic coverage / topic detailedness and overall quality of Web pages that were assigned by the evaluators. As shown in Fig. 5, we found out that topic coverage and topic detailedness of Web

Table 4. Spearman's rank correlation coefficient between original ranking (OR) or system ranking (SR) and user ranking (UR)

Query	topic coverage		topic detailedness	
	OR v.s UR	SR v.s. UR	OR v.s. UR	SR v.s. UR
support vector machine	0.6364	0.2970	0.5758	0.1758
Eric Clapton	0.2121	0.8061	-0.0061	0.6485
subprime lending	0.2121	0.4303	0.1152	0.6121
parkinson's disease	0.3697	0.3576	0.3333	0.4909
Hurricane Katrina	0.2242	0.4303	0.3576	0.5030
global warming carbon dioxide	0.2970	0.6848	0.3576	0.8182
iPod Zune comparison	0.1030	0.5879	0.1515	0.4424
ancient Olympics event	0.1394	0.5273	0.1030	0.6727
obesity causes	0.1152	0.6121	0.2242	0.5152
PageRank search engine optimization	-0.4182	-0.0424	-0.4182	0.0182
Avg.	0.1891	**0.4691**	0.1794	**0.4897**

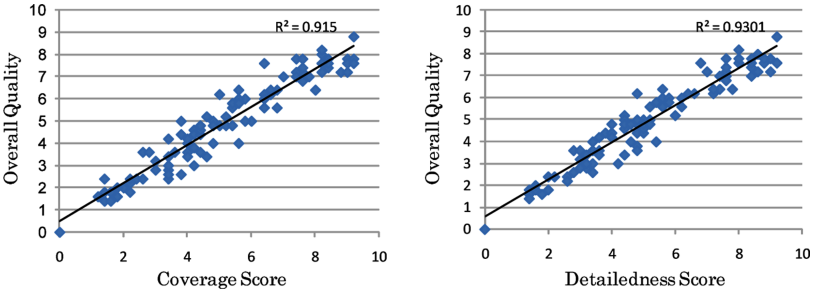

Fig. 5. Relationships between overall quality and topic coverage, topic detailedness

pages are strongly correlated to their overall quality. This means that topic coverage or topic detailedness are important factors for evaluating the quality of Web pages.

5 Conclusion and Future Work

In this paper, we introduced the notions of topic coverage and topic detailedness of Web pages which are important factors in evaluating their quality. Topic coverage and topic detailedness are calculated by using typical terms and special terms. We have proposed a method of measuring typicality and speciality of terms by utilizing structural features of Wikipedia. We also implemented a system that presents Web search results with the scores of topic coverage and topic detailedness of pages. The experimental results suggested that our proposed methods are effective in classifying topic terms and that automatic evaluation of Web page quality by topic coverage and topic detailedness has a positive correlation with a manual evaluation.

Our proposed methods using Wikipedia data still have some disadvantageous that need to be approached. One problem is the word sense disambiguation like in the example of "Java". We think that this problem can be solved by applying

disambiguation methods proposed, for example in [18][19]. Another problem is certain, inherent limitation of Wikipedia. Although Wikipedia contains a huge amount of content, it does not necessarily cover all the possible topics and the quality and scope of its articles may actually differ for different topics. However, currently Wikipedia is the largest, manually edited knowledge base available online. It is also frequently and promptly updated according to real world changes.

We focused only on two aspects of quality evaluations of Web pages, topic coverage and topic detailedness. Both are actually query-dependent and user-dependent quality measures. In our future work, we plan to combine the proposed methods with other query-independent quality measures such as a readability for more precisely evaluating quality of Web pages. We also intend to propose a system for re-ranking search results by user's knowledge level about a search query or its domain.

Acknowledgments. This work was supported in part by the following projects and institutions: Grants-in-Aid for Scientific Research (Nos. 18049041 and 18049073) from MEXT of Japan, a MEXT project entitled "Software Technologies for Search and Integration across Heterogeneous- Media Archives," a Kyoto University GCOE Program entitled "Informatics Education and Research for Knowledge- Circulating Society," and the National Institute of Information and Communications Technology.

References

1. Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Tezuka, T., Oyama, S., Tanaka, K.: Trustworthiness analysis of web search results. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 38–49. Springer, Heidelberg (2007)
2. Giles, J.: Internet encyclopedia go head to head. Nature 438 (2005)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30, 107–117
4. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (1999)
5. Haveliwala, T.H.: Topic-sensitive pagerank. In: WWW 2002: Proceedings of the 11th international conference on World Wide Web, pp. 517–526. ACM, New York (2002)
6. Cho, J., Roy, S., Adams, R.E.: Page quality: in search of an unbiased web ranking. In: SIGMOD 2005: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 551–562. ACM, New York (2005)
7. Yanbe, Y., Jatowt, A., Nakamura, S., Tanaka, K.: Can social bookmarking enhance search in the web? In: JCDL 2007: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 107–116. ACM, New York (2007)
8. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: WWW 2007: Proceedings of the 16th international conference on World Wide Web, pp. 501–510. ACM, New York (2007)

9. Amento, B., Terveen, L., Hill, W.: Does "authority" mean quality? predicting expert quality ratings of web documents. In: SIGIR 2000: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 296–303. ACM, New York (2000)
10. Ivory, M.Y., Hearst, M.A.: Statistical profiles of highly-rated web sites. In: CHI 2002: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 367–374. ACM, New York (2002)
11. Mandl, T.: Implementation and evaluation of a quality-based search engine. In: HYPERTEXT 2006: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp. 73–84. ACM, New York (2006)
12. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1992)
13. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. In: Proceedings of the 16th international conference on WWW. ACM, New York (2007)
14. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE TKDE 19(3) (2007)
15. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: Proceedings of National Conference for Artificial Intelligence (2006)
16. Milne, D., Medelyan, O., Witten, I.H.: Mining domain-specific thesauri from wikipedia: A case study. In: International Conference on Web Intelligence (2006)
17. Erdmann, M., Nakayama, K., Hara, T., Nishio, S.: An approach for extracting bilingual terminology from wikipedia. In: Haritsa, J.R., Kotagiri, R., Pudi, V. (eds.) DASFAA 2008. LNCS, vol. 4947, pp. 380–392. Springer, Heidelberg (2008)
18. Milne, D.N., Witten, I.H., Nichols, D.M.: A knowledge-based search engine powered by wikipedia. In: Proceedings of the sixteenth ACM conference on CIKM. ACM, New York (2007)
19. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on CIKM. ACM, New York (2007)
20. Bennett, N.A., Qin He, K.P., Schatz, B.R.: Extracting noun phrases for all of medline. In: Proceedings of the American Medical Informatics Association (1999)
21. Klavans, J.L., Muresan, S.: Definder: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In: Proceeding of the American Medical Informatics Association (2000)
22. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G.: Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth IJCAI. Morgan Kaufmann Publishers Inc., San Francisco (1999)
23. Liu, B., Chin, C.W., Ng, H.T.: Mining topic-specific concepts and definitions on the web. In: Proceedings of the 12th international conference on WWW. ACM, New York (2003)