Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I,
IMECS 2014, March 12 - 14, 2014, Hong Kong

Comprehensive Web Search based on Sentiment Features

Katsutoshi Minami, Shoko Wakamiya, Noriaki Hata, Yukiko Kawai, Tadahiko Kumamoto, Jianwei Zhang, and Yuhki Shiraishi

***Abstract*—We have developed a novel system for comprehensive Web search considered the diversity of sentiments. Recently, lots of studies and services based on sentiment analysis have been conducted, because it is still difficult to search and summarize information satisfying users' needs by text analysis only. In this paper, we propose a method to extract multidimensional sentiments of Web pages on a topic such as "Happy⇔Sad," "Glad⇔Angry," and "Peaceful⇔Strained." Specifically, the system can extract common sentiments for a topic and also retrieve and present various Web pages based on comprehensive sentiments. In the experiment, we evaluate our proposed method using the developed system.**

***Index Terms*—Web search, diverse sentiment analysis, visualization**

I. Introduction

General search engines such as Google and Yahoo! have become main tools to obtain information from the Web. They typically provide a list of search results for a given query, ranked by relevancy and popularity between a page and the query. Generally, users only look through Web pages in the top-ranked search results and may often find the information satisfying their search needs. However, current search engines sometimes fail to return user-desired pages at top ranks due to the diversity of users' search intentions.

Previous work have been devoted to improve search results in three main directions: (1) ranking pages in the initially retrieved list, (2) suggesting query expansion or reformulation for re-retrieving new pages, and (3) analyzing sentiments of pages for opinion mining. As for (1), many aspects have been utilized to achieve effective ranking of search results, such as query logs [1], authorship [2], social tags [3], re-finding [4], and multiple pairwise relationships between pages [5]. As for (2), query expansion or reformulation involves expanding or revising the search query to match additional or new pages by utilizing some technologies and information such as global analysis [6], and social annotation [7]. As for (3), sentiment analysis and opinion mining [8], [9] have attracted a lot of research interests, which study sentiments and its related concepts such as opinions and attitudes. Interestingly, the role of sentiment in information retrieval has been investigated in some researches [10]. Especially, researches on sentiment retrieval or opinion retrieval [11], [12], [13] aim to provide a general opinion search service. These are similar to traditional Web search in the way that both of them try to find pages relevant to a query. On the other hand, these are different from the latter in the way that sentiment retrieval needs further determinations whether the pages express opinions on the query topic and whether their polarities are positive or negative.

Manuscript received January 8, 2014; revised January 30, 2014.

K. Minami, N. Hata, and Y. Kawai are Kyoto Sangyo University, Japan, e-mail: i1258131,g1044895,kawai@cc.kyoto-su.ac.jp

S. Wakamiya is University of Hyogo, Japan, e-mail: s.wakamiya@gmail.com

T. Kumamoto is Chiba Institute of Technology, Japan, e-mail: kumamoto@net.it-chiba.ac.jp

J. Zhang and Y. Shiraishi are Tsukuba University of Technology, Japan, e-mail: zhangjw, yuhkis@a.tsukuba-tech.ac.jp

In this paper, we attempt to exploit comprehensive sentiment features of pages for effective Web search. For this, we adopt more diverse sentiments such as "Happy⇔Sad," "Glad⇔Angry," and "Peaceful⇔Strained," not restricted to positive-negative sentiment. In addition, users can specify not only the three sentiment features, but also strengths of the sentiments to rank and retrieve pages. These features are also utilized to perform a re-retrieval to obtain pages with various sentiments. Specially, we propose a system consisting of the following parts:

- constructing a sentiment dictionary that represents words and their sentiment values on three types of sentiments,
- extracting sentiments on a given query topic, i.e., the major sentiment tendency in results of search engines, and
- retrieving diverse articles in terms of sentiments.

Fig. I and Fig. I show examples of search results for a query topic 'child benefit.' Fig. I is an example of initial search results by google. The figure shows that the system provides sentiments with respect to the query topic. We also propose a diverse retrieval which can automatically find pages with minor sentiments tendency. Fig. I is an example of search results for same query by our proposed method, and it shows that each article is mapped by various sentiments. The query topic 'child benefit' in these examples is a law introduced in Japan about distributing social security payment to parents of children. The sentiments of initial search results on this topic are a little sad, a little angry, and a little strained (Fig. I), because most of articles in the initial search list introduce that many people decline this offer or parents may abuse this grant. When a user wants to read the article with other sentiments, he/she just clicks a point which is plotted on a sentiment graph.

The rest of this paper is structured as follows. Section 2 provides an overview of the system. Section 3 describes the sentiment calculation for search results. Section 4 explains how to retrieve comprehensive search using diverse sentiment features. Section 5 evaluates the effectiveness of our system. Section 6 reviews related work. Finally, we conclude the paper and discuss future work in Section 7.

ISBN: 978-988-19252-5-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

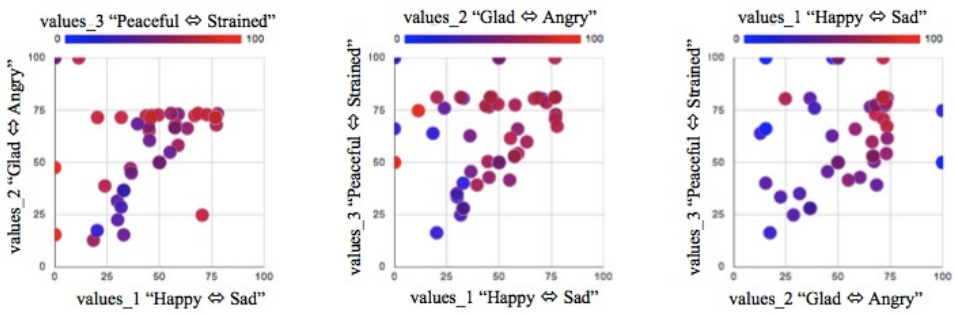

Fig. 1. Sentiment distribution of search results by google

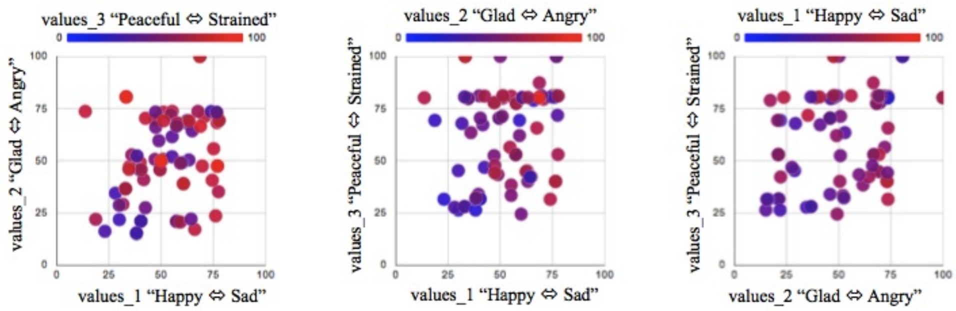

Fig. 2. Sentiment distribution of search results by our proposed system

II. System overview

Fig. II shows the overview of the proposed system. Given a query topic by a user, the system performs the following process:

1. Initial search results are returned by using Google! Web Search API [16]. Specially, titles and snippets of pages are obtained for sentiment analysis.
2. Both the initial search results and sentiments with respect to the query topic are presented to the user. Moreover, the diverse search results are presented to the user. The sentiments have three dimensions and are presented on a graph.

2a. The sentiments of each article from initial search results are calculated by using a sentiments dictionary, and the article is plotted on each sentiment graph.

2b. The sub-topic keywords are extracted by calculating $Tfidf$ from each article which has maximum of sentiment in all results, then
2b-1) the system retrieves by using the sub-topic keywords only, and extracts the articles which involved the user's input query from search results, and
2b-2) the system retrieves by using both the sub-topic keywords and the query.
The details of comprehensive search method are described in Section IV-A.

3. Finally, the initial search results and the comprehensive search results are presented to the user by the sentiment graphs. The user can obtain the sentiment tendency of the query from initial search results and each article based on various sentiments by using comprehensive search results.

III. Calculation of Sentiments with respect to Query Topic

A. Construction of Emotion Dictionary

We construct an emotion dictionary, in which each entry indicates the correspondence of a word and its sentiment features on three dimensions. The three-dimension sentiments are “Happy⇔Sad,” “Glad⇔Angry,” and “Peaceful⇔Strained,” that are formed based on a statistical analysis and a clustering analysis in our previous work [17]. A sample of the sentiment dictionary is shown in Table I. Sentiment feature $s(w)$ of a word w on each dimension is a value between 0 and 1. The values close to 1 mean the sentiments of the words are close to “Happy,” “Glad,” or “Peaceful,” while the values close to 0 mean the words' sentiments are close to “Sad,” “Angry,” or “Strained.” For example, the sentiment feature of the word ‘prize’ on “Happy⇔Sad” is 0.862, which means the word ‘prize’ conveys a “Happy” sentiment. The sentiment feature of the word ‘deception’ on “Glad⇔Angry” is 0.075, which means ‘deception’ conveys an “Angry” sentiment.

For each of the three dimensions, we set two opposite sets (OW_L and OW_R) of original sentiment words (Table II). The basic idea of emotion dictionary construction is that a word expressing a left sentiment on a dimension often occurs with

ISBN: 978-988-19252-5-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

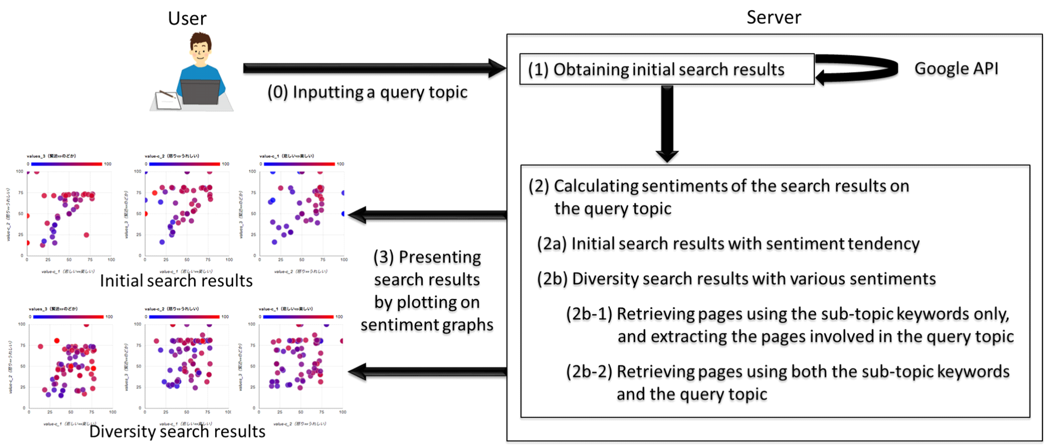

Fig. 3. Overview of comprehensive Web search system based on sentiment diversity

TABLE I
A SAMPLE OF THE SENTIMENT DICTIONARY

word w	$s(w)$ on Happy⇔Sad	$s(w)$ on Glad⇔Angry	$s(w)$ on Peaceful⇔Strained
'prize'	0.862	1.000	0.808
'cooking'	1.000	0.653	0.881
'deception'	0.245	0.075	0.297
'death'	0.013	0.028	0.000

TABLE II
ORIGINAL SENTIMENT WORDS FOR THREE DIMENSIONS

Dimension	Original sentiment words
Happy ⇔Sad	Happy, Enjoy, Enjoyment, Joy (OW_L) Sad, Grieve, Sadness, Sorrow (OW_R)
Glad ⇔Angry	Glad, Delightful, Delight (OW_L) Angry, Infuriate, Rage (OW_R)
Peaceful ⇔Strained	Peaceful, Mild, Primitive, Secure (OW_L) Tense, Eerie, Worry, Fear (OW_R)

the dimension's OW_L, but rarely occurs with its OW_R. For example, the word 'prize' expressing the sentiment "Happy" often occurs with the words "Happy," "Enjoy," "Enjoyment," "Joy," but rarely occurs with the words "Sad," "Grieve," "Sadness," "Sorrow." We compare the co-occurrence of each target word with the two sets of original sentiment words for each dimension by analyzing the news articles published by a Japanese newspaper YOMIURI ONLINE during 2002-2006.

First, for each dimension, we extract the set S of news articles including one or more original sentiment words in OW_L or OW_R. Then, for each news article, we count the numbers of the words that are included in OW_L and in OW_R. The news articles, in which there are more words included in OW_L than in OW_R, constitute the set S_L. Inversely, the news articles, in which there are more words included in OW_R than in OW_L, constitute the set S_R. N_L and N_R represent the numbers of the news articles in S_L and S_R, respectively. For each word w occurring in the set S, we count the number of news articles including w in S_L and mark it as $N_L(w)$. Similarly, we count and mark the number of news articles including w in S_R as $N_R(w)$.

Sentiment feature $s(w)$ of a word w is calculated as follows:

$$s(w) = \frac{P_L(w) * weight_L}{P_L(w) * weight_L + P_R(w) * weight_R}$$

where $weight_L = log_{10} N_L$ and $weight_R = log_{10} N_R$.

B. Sentiment Calculation for Individual Pages and Sentiment Summary of Search Results

The sentiment features of an individual page are calculated by looking up the sentiment feature of the words in the page from the emotion dictionary and averaging them [1]. In this way, a page has sentiment feature ranging from 0 to 1, since the sentiment feature of the words in the emotion dictionary range from 0 to 1. Considering the comprehensibility and the symmetry, the sentiment feature (x) of a page is further converted to a value (y) ranging from -3 to 3 by the formula: $(y = 6 * x - 3)$. When x is 1, 0.5, and 0, the corresponding y becomes 3, 0, and -3. The sentiment feature is 3, 2, 1, 0, -1, -2, -3 on a dimension, e.g., "Happy⇔Sad," corresponds to "Happy," "Relatively happy," "A little happy," "Neutral," "A little sad," "Relatively sad," and "Sad," respectively.

Sentiment feature on each dimension of search results in response to a query is averaged as the sentiment (polarity and strength) on that dimension with respect to the query. As shown in Fig. I and Fig. I, each axis has opposite sentiment polarities for each sentiment, thus represents the sentiment strengths as their absolute values, although the values of negative sentiments in the inner system are negative numbers.

IV. COMPREHENSIVE SEARCH BASED ON VARIOUS SENTIMENT FEATURES

A. Using Sub-topic Keywords with Correlation and Diverse Sentiments

After calculating sentiment tendency from initial search results, the sub-topic keywords with correlation and diverse

[1]Since the title and snippet of a page summarize the content of the page and their text is shorter than full page, the system actually calculates the sentiment feature using the text of the title and snippet for each page so as to shorten the response time.

sentiments are detected for retrieving and re-ranking the results. The system generates sentiment query vector $V_q = (v_{q1}, v_{q2}, v_{q3})$ using the sentiment feature on each dimension as its element. For each page in the initial search results list, sentiment page vector $V_p = (v_{p1}, v_{p2}, v_{p3})$ ($p = 1$, ..., N, where N is the number of initial search results) is determined, the elements of which are the sentiment feature on three dimensions of each page.

For obtaining pages reflecting various sentiments, we extract the sub-topics as follows:

1) The sentiments on the three dimensions with respect to the initial query (sentiment summary of initial search results) are represented as sentiment query vector $V_q = (v_{q1}, v_{q2}, v_{q3})$. Each page in the initial search results list is also represented as sentiment page vector $V_p = (v_{p1}, v_{p2}, v_{p3})$.
2) Each element is mapped to $y = x$ and $y = -x$ on each two dimensions $((v_{p1}, v_{p2}), (v_{p2}, v_{p3}), (v_{p3}, v_{p1}))$, and then each element of maximum and minimum are extracted using mapped values.
3) After extracting eight elements, by using extracted each element, the pages involving those elements are detected.
4) Keywords are extracted from the detected pages by using Yahoo! Term Extraction API [18]. The keywords whose scores are larger than 40 are determined as the candidate sub-topic keywords.
5) The system looks up the sentiment feature of each candidate sub-topic keywords from the sentiment dictionary, converts the sentiment feature to the scale ranging from -3 to 3, and forms sentiment sub-topic keywords vector $V_s = (v_{s1}, v_{s2}, v_{s3})$.

The keywords with high score are determined as the sub-topic keywords with diversity and is utilized to expand the initial query for the retrieval. After the retrieval with using sub-topic keywords, the system selects the pages which have the query keywords.

B. *Using Sub-topic Keywords and Query Keywords*

After detecting sub-topic keywords by following same process, the system retrieves pages using both sub-topic keywords and query topic keyword. The system does not need to determine the obtained results, because the results are related to the query keyword.

V. Experimental Evaluation

A. *Diverse Sentiment Analysis*

We have selected seven query topics for verifying the effect of proposed method, and then conduct a re-search as described in Section IV-A. Table IV shows the query topics and their extracted sub-topic keywords. We compare the variance of general search results (google) with proposed methods by using query and sub-topic keywords. The number of search results are 96 articles, respectively.

Table III shows the standard deviation of each search results for seven query topics. In the experiments, we observe that for almost query topics, both of proposed method become increasing than general search engine. Furthermore, the overall averages of each sentiment feature for all query topics become more diverse than general search engine.

TABLE IV
Query topics and opposite sub-queries (translated from Japanese)

	query	sub-topic keywords
1	'Liberal Democratic Party'	Microsoft, Opinions, we, MSN Sankei News, President, Kafesuta, Chiyoda-ku,Tokyo, Liberal Democratic Party, Topics by
2	'election'	Koizumi theater midst, Candidates database, Notification order, Liberal Democratic Party official, All domestic, Ballot box, Kawasaki, Japan's largest, Candidate poster board, Politician, Representative
3	'Special secret protection law'	Freedom of Information Act, Yosuke Isozaki, Much ado about nothing, Leakage, Night watchman, Good balance, Prime Minister aide, The ruling and opposition parties, Anonymous, Haver, Protection law
4	'tax increase'	Price, Consumption Tax Act, Public works, Tax increase issue, Monetary base, Bank of Japan, Asahi Shimbun published, Government budget, Revenue, Asset purchase policy, Market, Glossary
5	'TPP'	Commitment, Ratchet provisions, Agriculture, forestry and fisheries, Year agreement, WTO Doha Round, ISDS provisions, Trans-Pacific Partnership Agreement, Australian Government, Amari Minister, Economic partnership agreement
6	'Democratic Party'	Verbal abuse, Democratic-Party, Opinions, Turkmenistan, Thumbnail26, Republican Party, Mental abuse, Liberal Democratic Party, Toze, Two-party system, December 2013 Ohata Secretary-General briefing, President
7	'territorial issue'	Parties to each other, Partners, List of shops, Territorial sea base, Former Soviet Union, Border issue, Four islands conclusion of a peace treaty negotiations and northern, International Court of Justice, Claim, Ichiran-ya, U.S. Senate Senkaku problem, Mongolia

However, for example, for the query topics of 4 ('tax increase') and 5 ('TPP'), some sentiments of the general search results became higher than proposed methods. In that case, both query topics already have diversity because the trend has passed away.

B. *Novel Web Search Effect*

Fig. V-B and Fig. V-B show sentiment graphs. Fig. V-B is an example of initial google search results top 96 for a query topic 'territorial issue.' The figure shows that the system also provides the sentiments with respect to the query topic. We also propose comprehensive retrieval without user's input for automatically finding pages with minor sentiments tendency. Fig. V-B is an example of search results for same query by

TABLE III
RE-RANKING EFFECT

			Happy⇔Sad	Glad⇔Angry	Peaceful⇔Strained	
	query	search type	standard deviation	standard deviation	standard deviation	average
1	Liberal Democratic Party	(2b-1)	15.48	15.52	17.25	16.08
		(2b-2)	18.02	18.39	**20.39**	**18.93**
		normal	18.18	19.47	18.89	18.84
2	Election	(2b-1)	17.51	15.14	**16.63**	**16.43**
		(2b-2)	15.08	16.03	13.69	14.93
		normal	14.04	13.03	15.26	14.11
3	Special secret protection law	(2b-1)	10.58	15.41	14.99	13.66
		(2b-2)	16.73	21.43	**22.13**	**20.10**
		normal	16.31	17.47	16.97	16.91
4	tax increase	(2b-1)	18.03	12.37	12.78	14.39
		(2b-2)	14.76	16.17	14.93	15.28
		normal	17.62	19.37	**20.12**	**19.04**
5	TPP	(2b-1)	17.32	21.30	17.00	18.54
		(2b-2)	16.89	20.22	20.25	19.12
		normal	17.32	19.29	**21.46**	**19.36**
6	Democratic Party	(2b-1)	19.13	18.54	20.00	19.22
		(2b-2)	18.78	19.85	**22.02**	**20.22**
		normal	18.64	19.85	20.79	19.76
7	territorial issue	(2b-1)	**23.32**	11.84	8.63	14.60
		(2b-2)	20.21	19.02	15.01	**18.08**
		normal	18.24	18.22	11.95	16.14
	average	(2b-1)	17.34	15.73	15.32	16.13
		(2b-2)	17.21	**18.73**	18.35	**18.09**
		normal	17.19	18.10	17.92	17.74

using our proposed method, and it shows that each article is mapped on a graph by various sentiments.

The query topic 'territorial issue' in this example is in conflict with economic interests, such as resources between nations and territorial issue. The sentiments on this topic that initial search results reflect are a little sad, a little angry and a little strained (Fig. I), because most of articles in the initial search list introduce that territorial issue takes place in order to prioritize the interests of their own country. When a user wants to read the article with other sentiments, he/she just click points on the graph. For example, some articles arguing about 'territorial issue' which is a tense area of conflict between countries occurs express sad, angry, and strained sentiments.

VI. RELATED WORK

There are a number of studies on re-ranking search results considering various aspects. Zhuang et al. [1] proposed a Q-Rank method to refine the ranking of search results by constructing the query context from query logs. Bogers et al. [2] presented a passage-based approach to leverage information about the centrality of the document passages with respect to the initial search results list. Yan et al. [3] proposed a Query-Tag-Gap algorithm to re-rank search results based on the gap between search queries and social tags. Tyler et al. [4] utilized the prediction of re-finding (finding the pages that users have previously visited) to re-rank pages. Kang et al. [5] improved search results by using demographical contexts such as gender, age, and income.

Another research direction for improving retrieval accuracy is to re-retrieve new search results based on query expansion or reformulation. Xu et al. [6] utilized the hints such as query logs, snippets and search result documents from external search engines to expand the initial query. Lin et al. [7] proposed an automatic diagnosis of term mismatch to guide interactive query expansion or create conjunctive queries.

Similar to these researches we also aim to obtain a user-desired re-ranking and re-retrieval search results. Different from them we improve a personalized Web search using the sentiment features. The pages with sentiments similar to users' can be re-ranked to the top rank and the pages with opposite sentiments can be re-retrieved based on the extraction of opposite sub-queries.

On the other hand, sentiment analysis and opinion mining [8], [9] are one of the hottest research areas that extract sentiments (or sentiment, opinions, attitudes) from text such as movie reviews, book reviews, and product evaluations. Some researches have applied sentiment knowledge to information retrieval and its relevant research areas. Arapakis et al. [10] investigated the role of sentiment features in collaborative recommendation and their experimental results showed the sentiment features extracted from movie reviews were capable of enhancing recommendation effectiveness.

Specially, sentiment retrieval or opinion retrieval is a newly developed research subject, which requires documents to be retrieved and ranked according to opinions about a query topic. Eguchi et al. [11] proposed several sentiment retrieval models based on probabilistic language models, assuming that users both input query topics and specify sentiment polarity. Similar methods proposed in [12] and [13] unified topic relevance and opinion relevance respectively based on a quadratic combination and a linear combination combined topic-sentiment word pairs in a bipartite graph to effectively rank the documents. Opinion retrieval from UGC (User Generated Content) such as blogs [14] and Twitter [15] also yields comparable retrieval performance.

Different the above researches mainly focusing on review documents and positive-negative sentiments, we consider any Web pages and more diverse sentiments. As we showed in the experiments, there are Web pages that express both positive sentiment and negative sentiment in different dimensions.

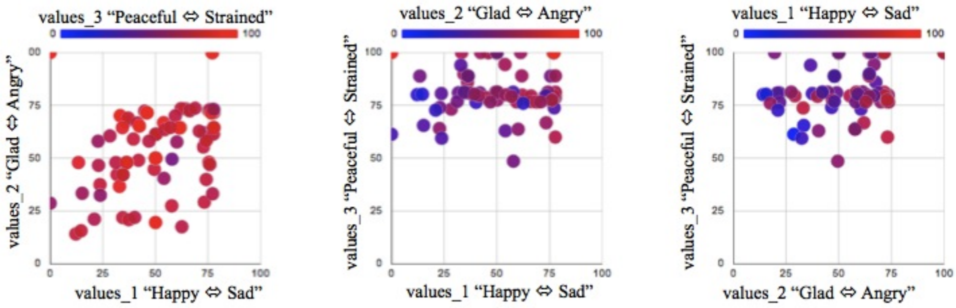

Fig. 4. Sentiment graphs plotted initial search results by google

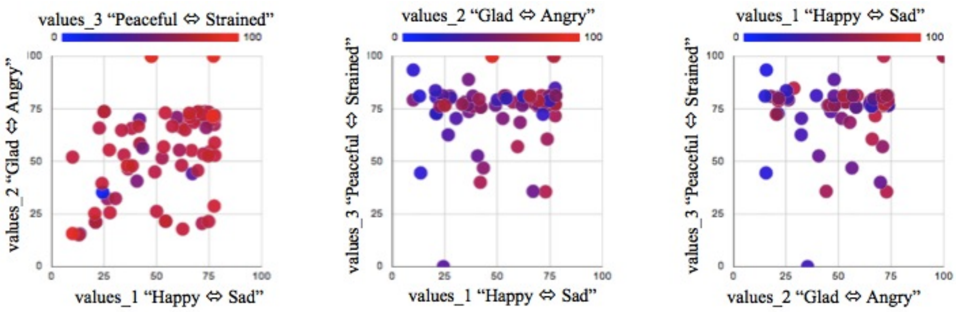

Fig. 5. Sentiment graphs plotted comprehensive search results based on our proposed method

VII. Conclusions

In this paper, we proposed a novel method for comprehensive Web search by calculating sentiment features of pages on a query topic. We developed a system which can present initial search results and comprehensive search results using sentiment graphs. By presenting initial search results, it could provide sentiment tendency of a query topic. From comprehensive search results, it could show exhaustive articles in terms of diverse sentiments. In future work, we plan to examine a method to improve scores of sentiment distribution by repeatedly re-searching. In addition, we will discuss how to visualize sentiment distribution of search results.

Acknowledgments

This work was partially supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE), JSPS KAKENHI Grant Number 24800007, and Competitive Educational and Research Projects in Tsukuba University of Technology.

References

[1] Z. Zhuang and S. Cucerzan. Re-ranking search results using query logs. In *CIKM 2006*, pp. 860-861, 2006.
[2] T. Bogers and A. Bosch. Authoritative re-ranking of search results. In *ECIR 2006*, pp. 519-522, 2006.
[3] J. Yan, N. Liu, E. Q. Chang, L. Ji and Z. Chen. Search result re-ranking based on gap between search queries and social tags. In *WWW 2009*, pp. 1197-1198, 2009.
[4] S. K. Tyler, J. Wang and Y. Zhang. Utilizing re-finding for personalized information retrieval. In *CIKM 2010*, pp. 1469-1472, 2010.
[5] C. Kang, X. Wang, J. Chen, C. Liao, Y. Chang, B. L. Tseng and Z. Zheng. Learning to re-rank Web search results with multiple pairwise features. In *WSDM 2011*, pp. 735-744, 2011.
[6] J. Xu and W. B. Croft. Query expansion using local and global document analysis.In *SIGIR 1996*, pp. 4-11, 1996.
[7] Y. Lin, H. Lin, S. Jin and Z. Ye: Social annotation in query expansion: a machine learning approach. In *SIGIR 2011*, pp. 405-414, 2011.
[8] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1-2, pp. 1-135, 2007.
[9] B. liu. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers 2012.
[10] I. Arapakis, J. M. Jose and P. D. Gray. Affective feedback: an investigation into the role of sentiments in the information seeking process.In *SIGIR 2008*, pp. 395-402, 2008.
[11] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *EMNLP 2006*, pp. 345-354, 2006.
[12] M. Zhang and X. Ye. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *SIGIR 2008*, pp. 411-418, 2008.
[13] X. Huang and W. B. Croft. A unified relevance model for opinion retrieval. In *CIKM 2009*, pp. 947-956, 2009.
[14] W. Zhang, C. T. Yu and W. Meng. Opinion retrieval from blogs. In *CIKM 2007*, pp. 831-840, 2007.
[15] Z. Luo, M. Osborne and T. Wang. Opinion retrieval in Twitter. In *ICWSM 2012*, 2012.
[16] google custom search.
`https://developers.google.com/custom-search/`
[17] J. Zhang, Y. Kawai, T. Kumamoto, S. Nakajima and Y. Shiraishi. Diverse sentiment comparison of news Websites over time. In *KES-AMSTA 2012*, pp. 434-443, 2012.
[18] Yahoo! Term Extraction.
`http://developer.yahoo.com/search/content/V1/termExtraction.html`