

Web Search for Real-World Events and Actions

Takeshi Okadome, Hajime Funai, Sho Ito, Junya Nakajima, and Koh Kakusho

Abstract—The method proposed in this paper searches for web pages using an event-related query consisting of a noun, verb, and genre term. It re-ranks web pages retrieved using a standard search engine on the basis of scores calculated from an expression consisting of weighted factors such as the frequency of query words. For the genres that are characterized by their genre terms, the method optimizes the weights of the expression. Furthermore, the method attempts to improve the scores provided of relevant pages by using machine learning techniques. In addition, some evaluations are provided to show the effectiveness of the method.

Index Terms—Event-related web pages, optimization, re-rank, web page search

I. INTRODUCTION

The recent development of the Internet and ubiquitous technologies has enabled us to generate writing contents by bots. For example, they can use the API provided by Twitter to tweet sentences (Makice, 2009), whereas some physical objects with sensors connected to the Internet post their “activities” to a blog (Maekawa et al., 2008). These technologies allow us to maintain a diary automatically. However, contents generated automatically by bots, usually consist of stereotyped expressions produced using a simple language model such as an N-gram model and thus they are monotonous and consist of insubstantial descriptions that do not reflect dynamic aspects of real-world situations. They stand out in sharp contrast to contents posted by (human) users, which are of a wide variety and contain informative descriptions that reflect real-world dynamics.

On the other hand, by analyzing web pages, many researches try to infer situations or states of the real world. analysis For example, Bollen et al. (2011) measured the public mood from tweets submitted to Twitter to predict the stock market. Using RFID tags, Perkowitz et al. (2004) formulated activity models by translating labeled activities, such as “cooking pasta”, into probabilistic collections of object terms, such as “pot”. Given this view of activity models as text translations, they defined activities from the web in an unsupervised manner. The fruitful results of these researches suggest that reusing web pages relevant to events and activities of daily life may provide us with an approach to automatically produce essays associated with daily events and activities.

This study aims at retrieving many web pages relevant to everyday events and activities that can be used as subject matter for the automatic generation of writing contents. In particular, this paper describes a web search method whose query is assumed to be a “verb” expressing a daily event or

activity and a “noun” representing an everyday physical object. The method re-ranks web pages that are retrieved using a search engine on the basis of the scores calculated from an expression consisting of weighted factors such as the frequency of query words. For each of the genres, the method optimizes the weights of the expression.

In this article, we assume that by using a standard web search engine, we can retrieve pages associated with an event (or an activity) related to a physical object by including a verb representing the event and a noun representing the physical object. For example, the Google search engine retrieves many pages relevant to an event related to a physical object: ‘(someone) hit a ball’ when we input “hit ball” as a query to the engine.

Our search method uses a standard web search engine API for a query consisting of (a) a noun, a verb, and a genre term, or (b) a noun and a genre term. Then, it re-ranks the retrieved web pages on the basis of the scores calculated from an expression consisting of weighted factors such as the frequency of query words.

For each genre, the method optimizes the weights of the expression. It also tries to improve the scores of relevant pages by using machine learning techniques.

II. OUTLINES OF THE SEARCH METHOD

We assume that by using a standard web search engine, we can retrieve pages associated with an event (or an activity) related to a physical object by including a verb representing the event and a noun representing the physical object. For example, the Google search engine retrieves many pages relevant to an event related to a physical object: ‘(someone) hit a ball’ when we input “hit ball” as a query to the engine.

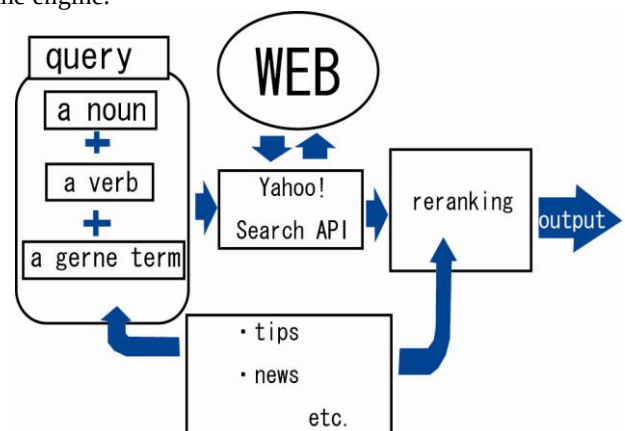


Fig. 1. Schematic representation of our method.

Our search method uses a standard web search engine API for a query consisting of (a) a noun, a verb, and a genre term, or (b) a noun and a genre term. Then, it re-ranks the retrieved web pages on the basis of the scores calculated

from an expression consisting of weighted factors such as the frequency of query words. For each genre, the method optimizes the weights of the expression. Figure 1 illustrates a schematic representation of the method.

III. GENRES AND QUERIES

This section describes some aspects of genres and queries that our search method deals with. In general, web pages are classified into genres such as tips, news, gourmet, shopping, etc. Thus, document genres are used to constrain the search space, with the intent of improving the search results (Rosso, 2005). This paper deals with the genres ‘tips’ (bits of knowledge), ‘how to,’ ‘news,’ and ‘short readings,’ because they have clear genre terms: “tips” for the genre ‘tips,’ “how to” for ‘how to,’ “news” for ‘news,’ and “short story” for ‘readings.’ These genre terms are widely used in daily web searches.

Our search method assumes a three-word query consisting of a noun, a verb, and a genre term. For a one-word query, a standard search engine returns pages covering a broad range of genres and topics and as a result we obtain too many irrelevant pages. On the other hand, a three or more word query reduces the number of web pages dramatically and we may miss relevant pages (Henziger, et al., 2003). therefore, our method assumes that a query consists of a noun, a verb, and a genre term for the genres ‘how to,’ ‘news,’ and ‘short readings.’ For some genres, however, a two-word plus a genre term query decreases the number of pages by too much. This happens to the genre ‘tips’ that we deal with. This is because many web pages about tips describe tips about a physical object, and not about an event associated with the object. Thus, for the genre ‘tips’ we use a two-word query, a noun and a genre term. Furthermore, for the genre ‘news,’ in addition to a three-word query, we try to use a two-word query, because we found that a three word query including verb reduced extremely the number of resulting pages in the genre.

IV. SCORING FOR WEB PAGES

Ranking pages in our method consists of the three steps: (1) retrieving web pages using a standard web search engine, (2) scoring the web pages, and (3) re-ranking the pages according to the scores.

A. Retrieving Web Pages

Our method uses one of the standard web search engines, the Yahoo search API. The Google search API in one search returns eight pages for a query, and for the same query, permits us to search for pages eight times or less. Thus, for the same query, we obtain at most 64 ages. On the other hand, the Yahoo search API returns 50 pages in one search and for the same query we can obtain 1,000 pages. For each query, we retrieve the top 100 web pages using the Yahoo search API.

B. Page Scoring Based on HTML

We score each of the 100 pages retrieved for a query. To score web pages, we define and use the following expression similar to that proposed by Lawrence and Giles

(1998).

$$\text{Score} = \omega_1 \cdot ES + \omega_2 \cdot FS + \frac{(\omega_3 - \min(DS, \omega_3))}{\frac{\omega_3}{\omega_1}} + \omega_4 \cdot SS + \omega_5 \cdot TS + \omega_6 \cdot LS + YS \quad (1)$$

where *ES*, *FS*, *DS*, *SS*, *TS*, *LS*, and *YS* are *score factors* and $w_1, \dots, w_6$ are weights that are optimized for each of the genres. The factor *YS* takes the score that the Yahoo search engine API returns. We describe below the other score factors and the optimization of the weights. Assume that we obtain a web page for a query.

ES: The factor *ES* represents the number of words in the query that occur inside the HTML body tag of the web page (we consider two-word genre terms as a single word.)

FS: The factor *FS* represents the number of occurrences of the query words in the page.

DS: The factor *DS* is the minimum distance between two words in the query that may reflect the degree of connection between the two words in the query.

SS: Let us, for example, take a query including “make” and “coffee” to retrieve web pages about “making coffee.” Then we may obtain a page including two successive sentences such as “I was drinking a cup of coffee. My children were making a castle out of building blocks.” Because the distance between “make” and “coffee” in the sentences is three and thus $DS = 3$, expression 1 produce a relatively high score for the page including sentences not relevant to “making coffee.” The factor *SS* reflects the idea that a web page may be more relevant if it contains a sentence that includes two words in a query.

TS: The factor *TS* represents whether the genre term in the query occurs in the title of the page.

LS: Many web pages in the ‘how to’ category favor the list style. Expression 1 captures this fact with the factor *LS*.

C. Optimization of the Weights

This study optimizes the weights of the factors in expression 1 for each of the genres. The optimization of the weights requires the manual assessment of web pages. That is, for each of the categories and for some queries, the web pages retrieved using the Yahoo search engine are classified into the following five classes according to their relevance to the queries: 1. Pages relevant to all the words in the queries, 2. Pages relevant to the noun and genre term in the queries, 3. Pages relevant to only the noun in the queries, 4. Pages non-relevant to the queries, and 5. No pages retrieved. This study defines two kinds of measures as a *relevance degree* for the five classes of pages. The first measure used for the classes ordered from 1 to 5 is (4, 2, 0, -2, -4), while the second measure used is (12, 4, 0, -2, -3).

Using the lists of the relevance degrees for each of the pages as a data set, we optimize the weights of expression 1 for each of the genres. The methods used for the optimization are simulated annealing (SA) and genetic algorithms (GAs). In the optimization we maximize the following *linear cumulative gain* (LCG) defined for an ordered list of search results returned for a query:

$$LCG = \sum_{i=1}^n ((n+1) - \text{rank}(i)) \cdot RS(i) \quad (2)$$

where n denotes the number of search results that are the top n pages returned by the Yahoo search engine. The function $rank(i)$ returns the rank of the i -th page based on its score calculated from expression 1, assuming the pages are numbered from 1 to n . $RS(i)$ is the relevance degree of the i -th page assessed manually.

In this study, for each of the four genres, we use the Yahoo search engine to retrieve the top 100 web pages for each of the 12 queries. In the queries, we choose nouns representing physical objects used in daily life such as “paper,” “tea,” and “chair.” In addition, from the most-frequently used verbs that are connected semantically (in a grammatical sense) to a noun, we select the verb so that a query consisting the nouns, the selected verbs, and a genre term many relevant pages contained in the top 100 pages.

D. Evaluation of re-ranking method

We evaluate the re-ranked search results by the leave-one-out cross-validation procedure. That is, for each of the genres, using expression 1 that has been optimized for each combination of 11 of the 12 queries, we re-rank the search results returned by the Yahoo search engine for the remaining query and calculate LCG for them. In the following, we describe the results of the optimization obtained by the genetic algorithm, which are better than those obtained by simulated annealing.

Figure 2 shows for the 12 queries for each of the genres, the rates of the number of queries for which LCG after re-ranking increases compared with that before re-ranking. In this figure, the results denoted by ‘genetic 1’ are obtained using the relevance degree measures: 4, 2, 0, -2, -4 and those denoted by ‘genetic 2’ 12, 4, 0, -2, -3. Also ‘News-2’ denotes the results for the two-word queries in the ‘news’ category and ‘News-3’ denotes those for the three-word queries. The results show that re-ranking using the optimized expression improves the ranks of the relevant pages for the genres ‘short story,’ ‘news-2,’ and ‘tips’. In particular, for the genre ‘tips,’ LCG increases for all 12 queries. Table 1 shows the average values of the optimized weights for the relevance degree measures: 4, 2, 0, -2, -4. We can see that the weights vary among the genres, indicating the contribution of the optimization, except those of *ES* that are relatively identical for all categories. Note that the weights of *DS* and *SS* for the genres ‘news-2’ and ‘tips’ are zero because the queries do not include verbs that affect *DS* and *SS*.

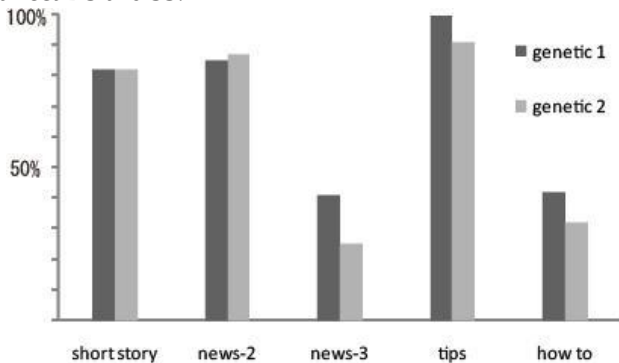


Fig. 2. The rate of the number of queries for which LCG after re-ranking increases compared with that before re-ranking (for the 12 queries for each of the genres).

V. FURTHER IMPROVEMENT BY ML

As described in the previous section, the re-ranking by optimizing the weight of the expression, improves the values of the scores of the relevant pages. Some pages with high-relevance degree, however, are ranked low. This is because, if the rank of a page with a low-relevance degree decreases due to the rank returned by a web search engine, the score also improves.

In this study, we attempt to further improve the ranking scores by applying machine learning techniques. That is, we first build classifiers that classify retrieved pages into two classes: relevant or non-relevant. Then, we add a new score factor to expression 1 that reflects the classification result of a page. For pages with high relevance degree, but low ranking, the new score expression with the classification factor is expected to increase their ranking such that they attain the first rank.

A. Gaining Scores

Supervised learning techniques permit us to build classifiers classify a retrieved page into two classes: relevant or non-relevant. Among the web pages classified into the five classes described before: (1) the pages relevant to all the words in the queries and (2) those relevant to the noun and the genre term in the queries are labeled as ‘relevant.’ The other pages are labeled as ‘non-relevant.’ In this study we use Weka 3 for supervised learning and from the learning algorithms, we select decision trees, neural networks, and support vector machines (SVN) with Gaussian and polynomial kernels. The classifiers constructed by learning from the data permit us to classify the web pages into ‘relevant’ and ‘non-relevant.’ Using the expression with the classification result of a page as a new score factor, we recalculate the score for the page.

B. Evaluation of re-ranking using ML techniques

For each of the genres, using the data of the 1,100 web pages for 11 out of the 12 queries, we construct the classifiers and test 100 web pages for the remaining query. The results show that the accuracy of the classification of the learning algorithms differ for the genres and on average, the SVM classifiers are the best among algorithms we used at classifying pages. Because only 33 % of all pages labeled manually are classified into the relevant class, the classifiers based on the training data tend to be biased towards ‘non-relevant.’ In particular, the classifiers constructed by SVMs are biased and hence their classification accuracies may not be reliable. On the other hand, the classification results of the neural network seem to be relatively reasonable. The re-ranking results using the learning algorithms, which are described in this section, reflect this conjecture.

Focusing on the re-ranking differences occurring when using the data forms, we first describe re-ranking performances based on the results for the neural network.

For the 12 queries for each of the genres, let N_i/N_d be the number of queries for which LCG after re-ranking using the expression with the classification factor increases/decreases compared with that without the classification factor. Figure 8 indicates $N_i - N_d$ for each of the genres. This figure reveals that the re-ranking method using the machine learning techniques improves the page rankings of the relevant pages.

In particular, using the boosted neural network we improve the rankings of the relevant pages for all queries.

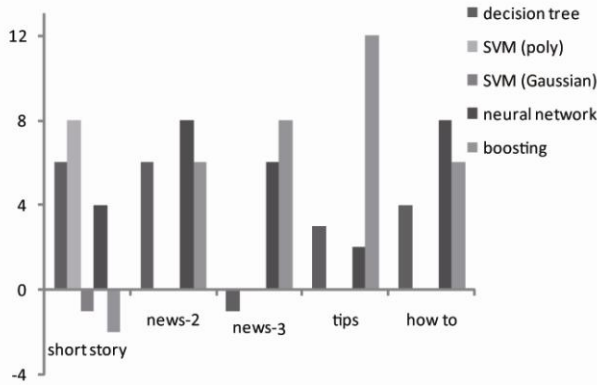


Fig. 3. $N_i - N_d$ for each of the learning algorithms, where N_i (N_d) is the number of queries for which LCG, after re-ranking using the expression with the classification factor increases (decreases) compared with that without the classification factor.

VI. DISCUSSION

Note that the re-ranking results for two-word queries are on the whole better than those for three-word queries. This is because some three (or more word) queries reduce dramatically the number of web pages and we may miss relevant pages (Henziger, *et al.*, 2003).

Table 1 shows that for the genre ‘short story,’ the optimized weight of the factor *SS* is larger than the weights of the other factors. This result implies that a short story that describes an event related to a physical object often includes a sentence in which both the verb and noun denote the event and object, respectively.

Table 1 also shows that for the genre ‘how to,’ the optimized weight of the factor *LS* is larger than The weights of the other factors. As expected, this result reflects the fact that many web pages about ‘how to’ favor the list style.

When using three-word queries in the ‘news’ genre, the weight of *FS* is zero. Furthermore, the weight of *TS* is smaller than the weights for the other genres using three-word queries. This may be due to the following reasons.

- Many news pages about an event or a physical object describe in detail what the event or the physical object is or how things led to the event. Thus, the word itself that denotes the event or the object may occur with relatively low frequency in the pages.
- A larger weight of *FS* or *TS* increases the rank of pages in the other genres because as shown in Table 1, the weights of *FS* and *TS* are large for the genres.

TABLE 1: OPTIMIZED WEIGHTS FOR THE GENRES.

	<i>ES</i>	<i>FS</i>	<i>DS</i>	<i>SS</i>	<i>TS</i>	<i>LS</i>
how to	12	20	47	50	62	16
Short story	11	24	50	74	44	0
news-3	20	0	45	17	10	0
news-4	14	31	0	0	10	2
tips	12	30	0	0	0	0

VII. RELATED WORK

Many studies have attempted to improve the rankings of results obtained by a search engine, by classifying web pages and/or using machine learning techniques. This section summarizes recent publications related to topics covered in this paper.

Bennett, Svore, and Dumais (2010) present a simple framework for classification-enhanced ranking that uses clicks in combination with the classification of web pages to derive a class distribution for a query. They then go on to define a variety of features that capture the match between the class distributions of a web page and a query, the ambiguity of a query, and the coverage of a retrieved result relative to a query’s set of classes. Their experimental results demonstrate that a ranker learned with these features, significantly improves ranking over a competitive baseline. Furthermore, their methodology is agnostic with respect to the classification space and can be used to derive query classes for a variety of different taxonomies.

Asadi, *et al.* (2011) examine the feasibility of constructing web search test collections in a completely unsupervised manner, given only a large web corpus as input. Within their proposed framework, anchor text extracted from the web graph is treated as a pseudo query log from which pseudo queries are sampled. For each pseudo query, a set of relevant and non-relevant documents are selected using a variety of web specific features, including spam and aggregated anchor text weights.

VIII. CONCLUSION

This study assumes that a standard search engine can retrieve pages associated with an event (or an activity) related to a physical object by including a verb representing the event and a noun representing the physical object. On the basis of this assumption, the method proposed in this paper searches for web pages using an event-related query consisting of a noun, verb, and genre term. The method re-ranks web pages that are retrieved using a search engine API on the basis of the scores calculated from an expression consisting of weighted factors such as the frequency of query words. For each of the genres that are characterized by their genre terms, it optimizes the weights of the expression.

ACKNOWLEDGMENTS

This research was partly supported by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan, Grant-in-Aid for Scientific Research (B), 23300073, 2011.

REFERENCES

- [1] N. Asadi, D. Metzler, T. Elsayed, and J. Lin, “Pseudo test collections for learning web search ranking functions,” in *Proc. of the 34th Annual SIGIR International Conference (SIGIR2011)*, pp.1073-1081, 2011.
- [2] N. Bennett, K. Svore, and S. T. Dumais, “Classification-enhanced ranking,” in *the Proceedings of the 19th International World Wide Web Conference (WWW2010)*, pp.111-120, 2010.
- [3] J. Bollen, H. Hao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, pp. 1-8, 2011.

- [4] D. Clark, "Shopbots become agents for business change," *IEEE Computer*, vol. 33, no. 2, 18-21, 2000.
- [5] M. Henzinger, B. W. Chang, B. Milch, and S. Brin, "Query-free news search," in *the Proceedings of the 12th International World Wide Web Conference (WWW2003)*, pp.1-10, 2003.
- [6] S. Lawrence and C. L. Giles, "Inquirus, the NECI meta search engine," in *the Proceedings of the 7th International World Wide Web Conference (WWW1998)*, pp.95-105, 1998.
- [7] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Kamei, Y. Sakurai, and T. Okadome, "Object-Blog System for Environment-Generated Content," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 20-27, 2008.
- [8] T. Makekawa, Y. Yanagisawa, Y. Sakurai, Y. Kishino, K. Kamei, and T. Okadome, "Web Search for Daily Living," in *the Proceedings of the 18th International World Wide Web Conference (WWW2009)*, pp.19-23, 2009.
- [9] K. Makice, *Twitter API: Up and Running: Learn How to Build Applications with the Twitter API*. O'Reilly Media, 2009.
- [10] M. Perkowitz, M. Philipose, D. J. Patterson, and K. Fishkin, "Mining models of human activities from the web," in *Proceedings of The Thirteenth International World Wide Web Conference (WWW2004)*, pp.573--582, 2004.
- [11] M. A. Rosso, "Using Genre to Improve Web Search," Ph.D. dissertation submitted to the faculty of the University of North Carolina, 2005.