Analysis of User Feedback Cost for Document Similarity Judgment

Minghuang Chen[*1] Seiji Yamada[*2] Yasufumi Takama[*1]

[*1]Tokyo Metropolitan University [*2]National Institute of Informatics

Abstract

This paper investigates the behavior of users judging the similarity of documents from the viewpoint of user feedback cost, in particular judgment time and accuracy. An experiment is conducted, in which 21 test participants were asked to judge the similarity of documents. As the clue for the judgment, 3 types of information: original text, snippet, and term, are mutually provided. The judgment accuracy and judgment time are analyzed using analysis of variance (ANOVA) and multiple comparison tests to examine the difference of snippet, term and text. The result shows that displaying term is the best in terms of time cost, whereas the judgment accuracy when a snippet is provided is improved with experience. The obtained result will contribute to the design of interfaces that can minimize the user's feedback cost.

1. Introduction

This paper investigates the behavior of users judging the similarity of documents from the viewpoint of user feedback cost, in particular judgment time and accuracy. Recent growth of the Web has brought us huge volume of information, which has already exceeded human capacity of information processing. In order to make use of available information, collaboration between human and computer systems is important. For example, the effectiveness of a Web search engine is not only determined by its retrieving mechanism such as ranking algorithm and crawlers, but also by the design of interaction with users, such as the page design of retrieved results.

One of the typical and promising approaches for realizing the collaboration between human and computer systems is to obtain feedback from users. Relevance feedback [1] obtains the result of user's relevance judgment of documents as the feedback for improving the retrieval performance. Constrained clustering [2] introduces two kinds of constraints: must-link and cannot-link into clustering process, and those constraints are supposed to be provided as user feedback.

When obtaining feedback from users, the workload of users providing feedback should be considered. Although much feedback information improves the effectiveness of computer systems, it forces heavy burden on users. In order to solve this tradeoff, the concept of Minimal User Feedback (MUF) [3] has been proposed, which aims at decreasing the cost of a user providing feedback information. One of the approaches for achieving MUF is to minimize the cost of generating each of feedback information. From this viewpoint, this paper focuses on the similarity judgment of documents. An experiment is conducted, in which test participants are asked to judge the similarity of two documents. Given a pair of news articles, a participant judges whether those articles relate with the same topic or not. As the clue for judging similarity, three kinds of information: original text, snippets, and terms, is mutually provided. As less work has been done for studying similarity judgment, it is not clear what terms or snippets are effective for the judgment. In this paper, we suppose that information identifying the difference and commonality of documents is effective. Therefore, common and specific terms / snippets are presented to test participants in a separate manner.

In order to examine the difference of snippet, text, and term in judgment time, analysis of variance (ANOVA) and multiple comparison tests are applied. The result shows participants viewing terms could judge the similarity of documents more quickly than viewing other conditions, whereas the improvement of accuracy with experience was observed when a snippet is presented.

2. Related works

The MUF employs two approaches: minimizing the quantity of feedback information and minimizing the

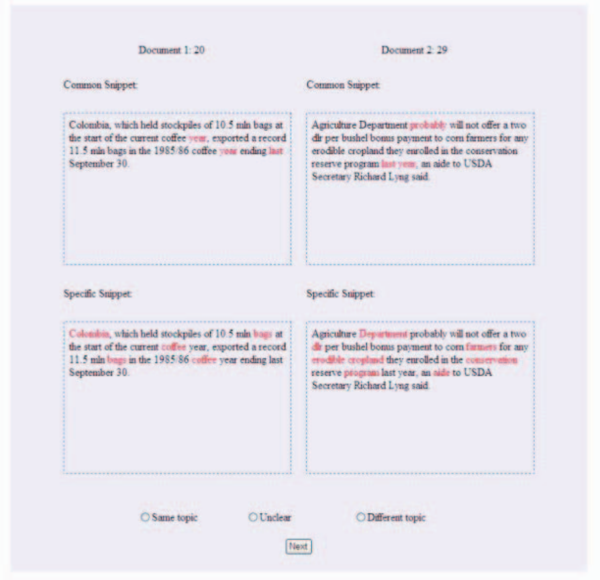
978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.116

cost of generating each of feedback information (i.e. relevance judgment for a single document). This paper addresses the latter approach, minimizing the cost of generating each of feedback information. A user usually generates feedback information by judging target objects. For example, a user judges the relevance of a document to a query in the case of document retrieval [1]. In order to provide must-links and cannot-links for constrained clustering [2], a user has to judge the similarity between target objects. Therefore, decreasing the cost of judgment is important.

Compared with similarity judgment, much work has been done for studying users behaviors in judging relevance of documents, which include users' viewing behaviors in search result pages and web pages [4, 5, 6], and study on the effect of snippet on relevance judgment [7, 8]. Chen et al compared accuracy of relevance judgment and judgment time between the condition of providing snippet and that of providing original text [8].

3. Outline of experiment

This paper investigates users' behaviors in similarity judgment. The task of test participants is to judge the similarity of two documents. Given a pair of documents, they are asked to judge whether those documents relate with the same topic or not.



Document 1: 20

Common Snippet:

Colombia, which held stockpiles of 10.5 mln bags at the start of the current coffee year, exported a record 11.5 mln bags in the 1985/86 coffee year ending last September 30.

Specific Snippet:

Colombia, which held stockpiles of 10.5 mln bags at the start of the current coffee year, exported a record 11.5 mln bags in the 1985/86 coffee year ending last September 30.

Document 2: 29

Common Snippet:

Agriculture Department probably will not offer a two dlr per bushel bonus payment to corn farmers for any erodible cropland they enrolled in the conservation reserve program last year, an aide to USDA Secretary Richard Lyng said.

Specific Snippet:

Agriculture Department probably will not offer a two dlr per bushel bonus payment to corn farmers for any erodible cropland they enrolled in the conservation reserve program last year, an aide to USDA Secretary Richard Lyng said.

Fig 1. A screenshot of the experiment system

For the experiment, we implement the experiment system that is written in VB language as ASP pages. Figure 1 shows the screenshot of the experiment system, which can be accessed with ordinary web browsers. In these pages, information about documents is arranged in two columns.

As noted in Sec. 3.2 and 3.3, two kinds of topic terms (snippet): common and specific terms (snippets) are presented. The common terms (snippet) are displayed in the upper part of the screen, and specific one is displayed in the lower part. Topic terms are highlighted with red when snippet is displayed.

3.1. Document set

The documents and topics are selected from Reuter Test Collection[1]. It includes 21578 documents with 135 topics. In the experiment, we prepared the document set by selecting a few topics and randomly picking up the corresponding documents. If the document of different topics is obviously different, test participants can judge the similarity of documents without carefully reading displayed information. Therefore, topics that are to be used in the experiment should relate with each other. Based on this consideration, we selected the following 3 topics: Coffee, Cocoa, and Corn. These topics are overlapping each other, i.e., several documents belong to two of those topics. Fourteen documents that belong to only one of those topics are collected from each topic, and total 42 documents are used in the experiment.

3.2. Extraction of topic terms

Terms that represent the topic of the document are supposed to work as a clue for judging similarity of documents. In particular, the terms indicating the difference and commonality between documents should be presented to a user. Based on this consideration, we classify the topic terms into common and specific terms, which are extracted with the following two steps.

Step 1: Extraction of topic terms from a document

Step 2: Extraction of common and specific terms

In step 1, given a set of documents D (42 documents used in the experiment), terms that have high TF-IDF values are extracted as topic terms. Among various definitions of TF and IDF, we employed the following equations.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \qquad (3.1)$$

$$IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|}, \qquad (3.2)$$

Where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, and the denominator of

[1] http://www.daviddlewis.com/resources/testcollections/

Eq. (3.1) is the sum of the number of occurrences of all terms in document d_j. The denominator of Eq. (3.2) shows the number of documents in which the term t_i appears. It should be noted that we calculate TFIDF score for only the terms appeared at least once in D. To be more exact, all of the terms in D are extracted and the TF-IDF values are calculated except the terms that are contained in a stop word list. In the experiment, we employed the stop word list available from Wikipedia[2].

In step 2, for a pair of documents that are to be compared, the topic terms that occur in both of the documents are selected as common terms, whereas the terms exclusively occur in either of the documents are selected as specific terms.

3.3. Snippet generation

One of the most important features of modern search engines is a snippet, which is a fragment of a document that represents its contents. In particular, the snippet generated based on the topic can help users to make a judgment easily on whether to read the corresponding documents or not. This merit is supposed to be valid for similarity judgment.

Based on the same consideration as noted in Sec. 3.2, two types of snippets, common and specific snippets, are employed in this paper. The snippets are generated by the following steps:

Step 1: Extraction of topic terms (Sec. 3.2).
Step 2: Score calculation for each sentence.
Step 3: Extraction of a set of sentences as a snippet.

In step 2, the score of a sentence is calculated based on the TF-IDF values of specific / common topic terms that are contained in the sentence.

In step 3, a set of sentences with the highest score is selected as a snippet. The snippet that consists of the sentences containing specific (common) terms is called specific (common) snippet.

4. Experimental results

In the experiment, 21 participants including graduate/undergraduate students and researchers in engineering field took part in the experiment. A participant is asked to judge 3 document pairs for each type of information (snippet, text, term). A pair of documents is generated randomly from the document set containing 42 documents.

When a user performs similarity judgment in an actual application, the judgment has to be repeated several times. Therefore, user's adaptability is one of important factors for evaluating the type of information provided for similarity judgment. As above-mentioned, a participant judged the similarity of documents 3 times for each type of information. In order to consider the participants' adaptability, we separately analyzed the results in the 1st and the 3rd trials.

Table 1 shows the experimental result of the 1st and the 3rd trial. The table contains average judgment time (AVG) on the second time scale, its standard deviation (STDEV), the number of correct answers, and mistakes.

Table 1. Experimental results

		AVG	STDEV	Correct	Mistake
1st trial	Text	73.31	40.00	12	9
	Snippet	40.07	27.62	11	10
	Term	36.41	27.80	10	11
3rd trial	Text	58.54	41.87	17	4
	Snippet	43.59	24.02	16	5
	Term	32.27	19.46	11	10

The difference of snippet, text, and term in judgment time of the 1st trial is analyzed using one-factor repeated measures analysis of variance (ANOVA). As a result, we found statistically significant differences in the mean judgment time among snippet, text, and term (F(2,40)=16.52, P=5.9E-06).

In the case of the 3rd trial, the assumption of equality of variance was rejected. Therefore, we conducted nonparametric test (Kruskal Wallis Test) and confirmed the difference is statistically significant ($\chi 2$=7.023, P=0.030).

As for the difference between the 1st and the 3rd trials, p-value of the 1st trial is much smaller than 3rd trial. We think that in the 1st trial, participants did not get used to the experiment including the type of information, which affected the variance.

Table 2. Multiple comparison test in the 1st trial

Level1	Level2	P-value (Tukey)	p-value (LSD)
Snippet	Text	0.0042**	0.0015**
Snippet	Term	0.9287	0.7152
Text	Term	0.0014**	0.0005**

In order to examine the effectiveness of each type of information, multiple comparison tests are conducted. In the case of the 1st trial, assumption of equality of variance could not be rejected. Therefore, Tukey's test and Fisher's LSD is used. Table 2 summarizes the result. In the paper, * and ** indicate the significant level

[2] http://en.wikipedia.org/wiki/Stop_words

of 5% and 1%, respectively. The result shows that the participants could judge the similarity of document using snippet and terms more quickly than reading original text.

As already noted, the assumption of equality of variance was rejected in the case of the 3rd trial. Therefore, we conducted nonparametric tests: Scheffe test and Stell-Dwass test, of which the results are shown in Table 3. In this case, only the difference between text and term is statistically significant. From both the results of the 1st and the 3rd trials, it is shown that providing terms is more effective in terms of the time cost of similarity judgment.

Table 3. Multiple comparison test in the 3rd trial

Level1	Level2	P-value (Scheffe)	p-value (Steel-Dwass)
Snippet	Text	0.5393	0.4264
Snippet	Term	0.3112	0.2206
Text	Term	0.0307*	0.0341*

A chi-square analysis on the number of correct answers and mistakes as shown in Table 1 are performed in order to investigate the effect of type of information on the accuracy of similarity judgment. Although we found no significant difference among 3 types of information in both of the 1st (χ2=0.382, P=0.826) and the 3rd trials (χ2=4.672, P=0.097), we can see the tendency that the difference in the 3rd trial is larger than the 1st trial. In particular, the judgment accuracy in the 3rd trial when snippet is provided gets improved from the 1st trial. However, the judgment accuracy when terms are provided is low both in the 1st and the 3rd trials.

We suppose this result indicates that snippets and original text are easier for the participants to adjust than terms. Additional experiment will be required to investigate whether or not the judgment accuracy with terms could be improved with more experience.

We also analyzed the behavior of the participants by an eye-tracking system, and found some interesting patterns of behavior [9]. In particular, it is observed that the behavior patterns are very different between snippet and term condition, which is considered to be connected with the above-mentioned results.

5. Conclusions

This paper investigates the behavior of users judging the similarity of documents from the viewpoint of user feedback cost, in particular judgment time and accuracy. The aim of the investigation is to obtain the hint for minimizing the cost of users judging similarity of documents, which is an essential task for users when performing interactive document clustering.

An experiment system was implemented, with using which 21 test participants were asked to judge the similarity of given pair of documents. As for the clue for the judgment, 3 types of information: original text, snippet, and term, are compared. The judgment accuracy and judgment time are analyzed by ANOVA and multiple comparison tests, and the result shows that presenting terms is the best in terms of time cost, whereas judgment accuracy when snippet is presented gets improved through experience. The relationship between these results and participants' typical behavior will be examined in more details in the future work.

Our future work include the design of interface that supports interactive document clustering based on constrained clustering method. The obtained result will contribute to realization of interface that can minimize the user's feedback cost.

References

[1] J. J. Rocchio, "Relevance feedback in information retrieval," In The SMART Retrieval System: Experiments in Automatic Document Processing, 1971.

[2] S. Basu, I. Davidson, and K. Wagstaff eds., Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapman & Hall, 2008.

[3] M. Okabe and S. Yamada, "Semi-supervised Query Expansion with Minimal Feedback," IEEE Trans. Knowledge and Data Engineering, Vol.19, No.11, pp.1585-1589, 2007.

[4] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay, "The influence of task and gender on search and evaluation behavior using Google," Information Processing and Management Vol. 42, No. 4, pp. 1123-1131, 2006.

[5] K. Rodden and X. Fu, "Exploring How Mouse Movements Relate to Eye Movements on Web Search Results Pages," SIGIR 2007 Workshop on Web Information Seeking and Interaction (WISI), pp.29-32, 2007.

[6] Q. Li, L. Sun, and J. Duan, "Web Page Viewing Behavior of Users: An Eye-Tracking Study," Int'l Conf. on Services Systems and Services Management (ICSSSM'05), pp. 244-249, 2005.

[7] E. Cutrell, Z. Guan, "An eye-tracking study of information usage in Web search: Variations in target position and contextual snippet length," CHI'07, pp. 407-416, 2007.

[8] M. Chen, S. Yamada, and Y. Takama, "An Investigation of Snippet Generation for Minimum User Feedback," 23rd Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2009), 2B2-1, 2009.

[9] M. Chen, S. Yamada, and Y. Takama, "Eye-tracking Analysis of User Behaviors in Document Similarity Judgment," 24th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2010), 2G2-OS9-3, 2010.