Towards Serving "Delicious" Information within Its Freshness Date

Hao Han
Kanagawa University, Japan
han@computer.org

Junxia Guo
Beijing University of Chemical Technology, China
gjxia@mail.buct.edu.cn

Takashi Nakayama
Kanagawa University, Japan
nakayama@info.kanagawa-u.ac.jp

Keizo Oyama
National Institute of Informatics and SOKENDAI (The Graduate University for Advanced Studies), Japan
oyama@nii.ac.jp

ABSTRACT

Like freshness date of food, Web information also has its "shelf life". In this paper, we exploratively study the reflection of shelf life of information in browsing behaviors. Our analysis shows that the satisfaction of browsing behavior is modified if the shelf life of information could be considered by search engines.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Algorithms, Experimentation, Performance

Keywords

Browsing Behavior; Information Retrieval; Search Engine; Shelf Life

1. INTRODUCTION

Web information has various properties. Like freshness date of food, information keeps "deliciousness" within its "shelf life", which is the recommended maximum time for use or consumption. However, the most-used search engines, such as Google and Bing, are mainly based on semantic similarity (between query keywords and target documents) and link analysis algorithms (e.g., PageRank). They cannot avoid the problem like "serving unsavory/out-of-date information past shelf life" since shelf life of information is not seriously considered. For example, search results responded to query "Kyoto sightseeing" may be referred even the information reflects the situation of 10 years ago. Meanwhile, to query "USB memory price", the information of the last year is mostly out of date.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742747..

In this paper, we present an on-going analysis of the browsing behaviors by using browsing log, accessed Web content and manually assessed query satisfaction as data sets. Extensive experiments show that time related features of information could act as an effective role in the modification of search engines.

2. QUERY SATISFACTION

We exploratively analyze the diverse features affecting the query satisfaction in the browsing behaviors of client users, and present the effectiveness of time-related features.

2.1 Data Set

Browsing log data and Yahoo! Chiebukuro data are used as our data set. Browsing log data is a log of webpage access collected for research and investigation. It contains 81,168,263 records of webpage access made by 24,498 panel users of different ages and genders. Since the browsing logs of different users are relatively independent, we separate them into independent query processes/sessions according to the session characteristics [2]. We selected Yahoo! Chiebukuro (Japanese Yahoo! Answers) webpages as the experimental data because they are well structured and richly contain time related information. After the data preprocessing, there were 120,182 query processes related to Yahoo! Chiebukuro.

2.2 Features of Satisfaction Level

The assessment of query satisfaction level could be considered a type of classification task and machine-learning techniques are employed. For selecting the most effective features used in the assessment, we use the non-text features, text-query features, and time-related features extracted from webpages and browsing log data.

According to our manual analysis and the existing researches, non-text features of a Chiebukuro webpage are selected to represent the basic attributes like "number of words in the question's text", "number of words in the best answer's text", "average number of words in the answers to a question", and "number of webpage viewers".

Text-query correlative features are selected based on the analysis of the abovementioned separated query processes. They are "relevance between query keywords and the text of the question and best answer", "relevance between query keywords and the text of question", "relevance between query keywords and the title of question", and "length of query keywords excluding postpositional particles". These relevance values are calculated by TF-IDF based on the morphological analysis of the concerned words.

Similarly, we select time-related features from different views like "average dwell time of browsing on webpage" and "duration from information generation (date when the question was resolved) to webpage browsing".

2.3 Experiment and Analysis

Three assessors were requested to independently label 1,500 records of training data with labeled satisfaction scores (1: satisfaction, -1: dissatisfaction, and 0: not sure). A moderate agreement (Fleiss' Kappa = 0.4901) is reached after removing the records labeled "not sure", and 1,486 valid records are left for further analysis and experiments. If different scores were labeled for one record, the major scores were selected as the final score.

Every learning algorithm tends to suit some problem types better than others. Thus, our experiments utilize Weka2 [3], among which, C4.5, SVM, Naive Bayes, Logistic, and AdaBoost are employed. Here, 6-fold cross-validation is adopted and we use different sets of features to observe the influence and effectiveness of the features.

Table 1 shows the results ***without*** the time-related features, and Table 2 shows the results ***with*** the time-related features. These results clearly explain the importance of the time-related features, especially in the C4.5 and AdaBoost models.

Table 1: Results Without Time-related Features

Classifier	Precision	Recall	F-Measure	ROC Area
Naive Bayes	0.6	0.623	0.606	0.687
Logistic	0.556	0.555	0.555	0.635
SVM	0.577	0.598	0.565	0.617
AdaBoost	0.552	0.6	0.57	0.652
C4.5	0.555	0.578	0.565	0.617

Table 2: Results With Time-related Features

Classifier	Precision	Recall	F-Measure	ROC Area
Naive Bayes	0.702	0.743	0.721	0.849
Logistic	0.697	0.685	0.691	0.782
SVM	0.729	0.745	0.727	0.764
AdaBoost	0.784	0.845	0.811	0.834
C4.5	0.776	0.835	0.803	0.858

3. SHELF LIFE

The assessors were also requested to label "shelf life" (within-shelf-life or out-of-date) since some of the "dissatisfaction" instances are caused by out-of-date information at the time of the query. Here, a record is identified as available information within its shelf life if two and more of the three assessors label it as available and satisfactory one.

Statistics of shelf life are made for out-of-date records in the categories "Politics and International Affairs", "Sports and Outdoors", "Children and Schools" and "Business and Economics", which are more time-sensitive and bring more dissatisfaction caused by shelf life in query process. For each out-of-date record, the interval months are calculated from information generation (production date) to webpage browsing (consumption date), and the statistical result is given as shown in Figure 1, which shows that the shelf lives of out-of-date information are mostly distributed within a range between 1 - 3 years.

Table 3 gives an analysis result of the frequencies of nouns and verbs (word stems) characteristically occurring in out-of-date information but seldom occurring in other information of the same category. Most of them are time-sensitive like "discount, popularity, annual rate", or are related to sudden incidents like "mafia (violent crime)", or reflect periodic/specified events like "(politics) voting, (specified) angling (event), (entrance) examination (questions)", which could be used for trend analysis [1].

Figure 1: Shelf Life

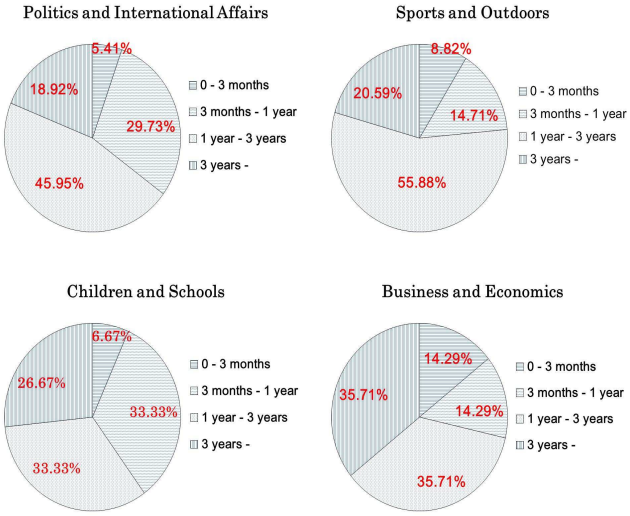

Table 3: Characteristic Words of Out-of-date Information

Category	Characteristic Words
Politics and International Affairs	voting, mafia ...
Sports and Outdoors	discount, angling ...
Children and Schools	examination, popularity ...
Business and Economics	exchange, annual rate ...

4. CONCLUSION AND FUTURE WORK

In this paper, we have proved that query satisfaction is visibly affected by time-related features, and presented the statistics of shelf life and characteristic words of out-of-date information.

As future work, we will conduct further analysis, such as calculating the value of "shelf life" of Web information, for further time-oriented optimization in search engines.

5. ACKNOWLEDGMENTS

This work was funded by the National Institute of Informatics and Kanagawa University under joint research grants.

6. REFERENCES

[1] S. Chelaru, I. S. Altingovde, S. Siersdorfer, and W. Nejdl. Analyzing, detecting, and exploiting sentiment in Web queries. *ACM Transactions on the Web*, 8(1), 2013.

[2] J. Guo, C. Gao, N. Xu, G. Lu, and H. Han. Analyzing query trails and satisfaction based on browsing behaviors. In *The Proceedings of the 10th Web Information System and Application Conference*, pages 107–112, 2013.

[3] Weka2. http://www.cs.waikato.ac.nz/ml/weka/.