ISSN 1346-5597

National Institute of Informatics

NII Technical Report

Web Content Summarization Using Social Bookmarking Service

Jaehui Park, Tomohiro Fukuhara, Ikki Ohmukai, and Hideaki Takeda

NII-2008-006E
Apr. 2008

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Web Content Summarization Using Social Bookmarking Service

Jaehui Park
School of Computer Science and Engineering
Seoul National University
Seoul 151-742, Republic of Korea
+82-2-880-5517
jaehui@europa.snu.ac.kr

Tomohiro Fukuhara
RACE (Research into Artifacts, Center for Engineering)
The University of Tokyo, Kashiwa
Chiba 277-8568, Japan
+81-4-7136-4275
fukuhara@race.u-tokyo.ac.jp

Ikki Ohmukai[1], Hideki Takeda[2]
[1]Digital Content and Media Science Reserch Division,
[2]Principles of Informatics Research Division
National Institute of Informatics, Tokyo
101-8430, Japan
+81-4-7136-4275
{i2k,takeda}@nii.ac.jp

ABSTRACT
In this paper, we propose a novel web content summarization method that creates a text summary by exploiting user feedback (comments, annotations etc.) in the social bookmarking service. We first analyze the feasibility to utilize user feedback in the summarization, and then demonstrate how the *social summary* which best represents the topics of the web content can be generated. Performance evaluations on our method are conducted by comparing its output summary with the manual summaries generated by human evaluators.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: Abstracting methods; H.3.3 [**Information Search and Retrieval**]: Information filtering, Selection process

General Terms
Experimentation.

Keywords
Social summarization, Social bookmarking service, user feedback

1. INTRODUCTION
Weblogs (Blogs) provide individual users with an easy way to publish online and others to comment on that. This characteristic of blogs makes them quite different from other internet contents. This could be used as a medium for social conversation. Permitting readers to participate the conversation, the blog entry can have a chance to aggregate related information from so-called social knowledge. Recently, many of internet services such as social bookmarking allow users to gather blog entries and to comment on it. As these services have matured and grown more popular, social knowledge which we can exploit will become richen and qualified.

In spite of blog entries containing comments, existing studies are largely focused on the post only. In this paper, we describe a summarization technique using user comments. First, we analyze comments of blog post in social bookmark service. The focus of our analysis is on the finding features for blog summarization from comments. Furthermore, we demonstrate how this summarization technique can generate that are of similar quality that those produced by human.

2. BACKGROUND
In this paper, we suggest an idea for summarization that is inspired by the recent emergence of so-called social services. By encouraging users to submit a note or a tag, they can express their views of the contents. The point of all this approach is based today's active user role in web environment. There are so many emerging services that allow users to play more active roles, eventually enriching or enhancing the original content.

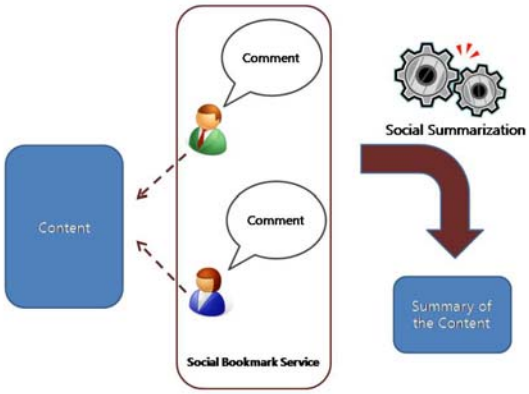

Figure 1. The Concept of Social Summarization

Based on user's active role, enriched data from the social services such as social bookmark services can be exploited for social summarization. Social summarization focuses on human's act toward information. Figure 1 shows the concept of social summarization.

3. PREVIOUS WORK
Much work has been done on Weblog Contents summarization. In this section we give a brief overview of

Nevertheless, very few studies on blog comments and blog post summarization have been reported [1][2]. The most recent one is

[2]. In this approach aim to extract representative sentences from a blog post that best represent the topics discussed among its comments. This paper shows information hidden in comments.

4. ANALYSIS OF COMMENT

Social bookmark services run as a tool for gathering various users' comments. Here we choose representative for this: Del.icio.us and Digg [3][4]. Del.icio.us is a social bookmark service for storing and sharing web bookmarks. This is one the most popular social bookmark services. Users can tag a bookmark or put a little note on it. Digg is a community-based news article site. It combines a social bookmark service and blogging service. Once one user post a bookmark, other users can simply rate (called diggs) or comment on the article.

We examined user note in Del.icio.us and comment in Digg. In spite that both have characteristics of their own service, we just focused on the contents of a bookmark and comments on it.

Table 1. Statistics of Experiment Data

Parameters	Value
Number of participant	10
Number of Articles	70
Number of Comments	2504
Data Type	Text, Video, Image

Firstly, for understanding patterns in comments, we collect 10 participants who are graduate students majoring in computer science. They have read comments on a number of articles during a week, and 70 articles and 2504 comments were gathered. Table 1 shows the statistics of experiment data. Gathered comments are manually categorized into 5 groups:

4.1 Categories

4.1.1 Summary

A main summarization feature for this work. Mostly, a collection of important sentences or sentence fragments from the original article are classified into this category. And paraphrase for the source content to provide a more concise representation is included.

4.1.2 Additional Information

Related information from the external sources such as related links, quotations, comparisons with another post, and so on.

4.1.3 Impression

An expression for describing user's sentiment when he sees the article such as "Awesome", "nice background design". Most of them includes adjective phrase.

4.1.4 Opinion

An explicit opinion against the article expressing a judgment or an assessment. This information could be used for enriching the original article in social ways. In a sense, getting the point is complicate work even for human.

4.1.5 Etc.

Spam, slang, non-English, joke, sarcasm, all including irrelevant information

Categorizing comments, we found some patterns according to social bookmark services: Del.icio.us and Digg, Figure 2 shows the different distribution of comment categories between them.

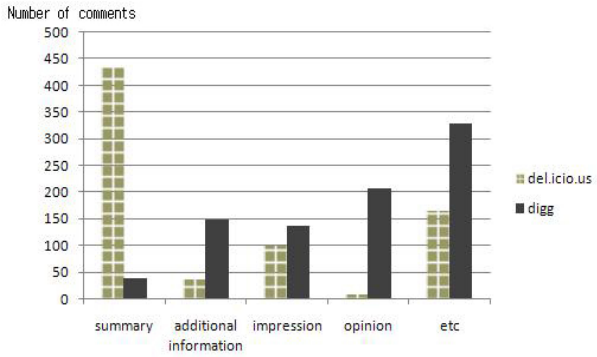

Figure 2. Number of comments in Del.icio.us and Digg

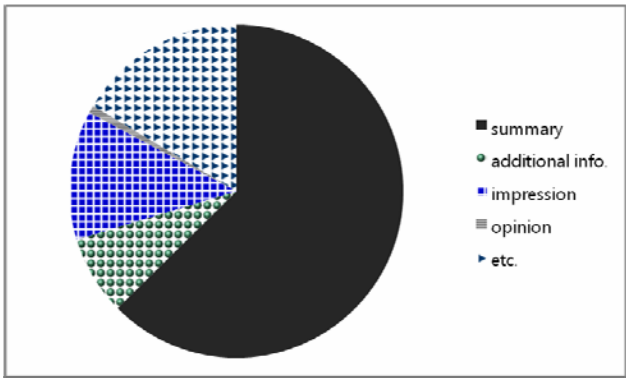

Figure 3. Pie Chart for Del.icio.us

Like Figure 3, in Del.icio.us, over 62% of comments are classified into category 'Summary'. In 14.6% of Etc, there are many comments written in other languages like Japanese. If we could understand other languages, the portion for 'Summary' might be bigger. This result shows that users of Del.icio.us have a tendency to summarize the contents of the bookmark. In user page there are all the bookmarks he/she made and user notes for each. When users create a bookmark, they put a user note on it. This page can be exposed to other users, but cannot be altered or added by them. This means that users have no place to discuss with other users, but more for their own uses such as summarizing original post concisely. Based on this analysis, comments in Del.icio.us are thought to be suitable for summarization.

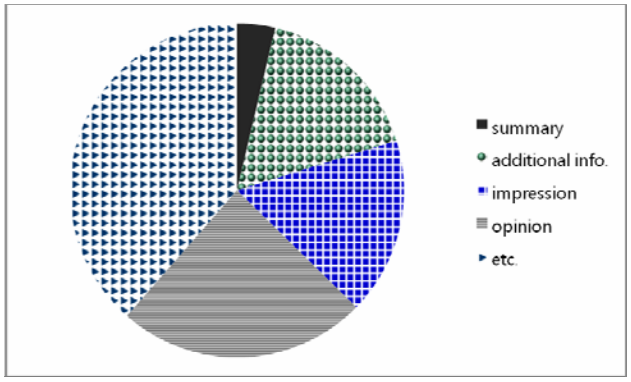

Figure 4. Pie Chart for Digg

Like Figure 4, in Digg, about 40% of their comments were cannot be processed. Most of them were jokes and sarcasm. However, comments classified into category 'Opinion' occupy over 25%. This includes useful information supports the original one. This is because bookmarked (called dugg) web contents exposed in the page which every user can add comments. Everyone express their thought by comments. Some comments like a shrewd criticism or a detail feedback are regarded as a social knowledge of good quality.

In Digg we found twice more the number of comments than in Del.icio.us. However, because of good quality of summaries in del.icio.us, we made use of comments in del.icio.us for our experiments. Based on our comment analysis, we expect that a comment list in the social bookmark services could be very useful for blog summarization.

5. EXPERIMENTS

We investigate how comments in social bookmark services are contribute to the blog summarization. The comment list are collected from Del.icio.us Today's popular items. On the average, there are 25 comments for one bookmark and over 62% of comments are good summaries for the bookmarked page. The final goal of this experiment is finding good set of sentences from comment list on the assumption that most of comments in del.icio.us have good summary.

Firstly, we extract words from comment list based on statistical natural language processing algorithm. In a collection of comment, we apply conventional tf-idf weighting algorithm to each word in a comment. The weight increases proportionally to the number of times a word appears in the comment but is offset by the frequency of the word in the collection. This weighting scheme is often used by information retrieval field such as search engine [5]. But conventional tf-idf algorithm work at poor performance in our experiment. This is because comments in the same collection mostly have a same set of topics. The rank of a word relevant to a small set of topics falls same as a rank of the word can be used when talking about almost any topic or any context.

Hence, we employ POS (Part of speech) tagger developed by The Stanford Natural Language Processing Group [6]. This tool reads text and assign parts of speech to each word, such as noun, verb, adjective, etc. Based on the comment analysis before, we determine that noun, verb, adjective parts will be gained more weights than others. This selection policy is added into our ranking function. And the English stop word dictionary is added to filter out word has no meaning such as 'is', 'the' and so on [7]. These improve the original word frequency algorithm performance notably.

Using this customized tf-idf algorithm we calculate weight of each word appeared in comment list. After that, we sum this weight up for constructing one complete sentence. Here we see one of the examples.

Figure 5. Del.icio.us example page

Figure 5 shows comment list (user note in Del.icio.us) page for bookmark which links to blog post titled "Why I'm excited about the Google Social Graph API – Bokardo". In this comment list, there are all 39 sentences. Sentences are classified into categories as follows, Table 2.

Table 2. Comments classfied into 5 categories

Summary	20
Additional Information	3
Impression	2
Opinion	5
Etc.	9

Knowing that these examples have about 50% summarization features, we run our algorithm. Among the weighted result sentences, we choose top-k sentences in Table 3.

Table 3. Top 10 sentences extracted and weight

category	sentence	weight
summary	The Google Social Graph API is a new programming API that allows developers to expose social relationships embedded in web sites	37.5957
summary	What if you want to combine your Amazon book history with your friends lists at Facebook so that you can see what your friends are reading	25.7348

summary	The Social Graph API helps solve this "silos of information" problem by allowing people to write software that understands who your friends are	23.4909
additional information	The truth is that Facebook Amazon or even Twitter never had a good glimpse of my true social network anyway	23.1096
summary	Do you ever feel like your personal information is spread across the web in a whole bunch of separate places	22.3561
Etc.	I never gave Facebook my email list they don't know anything about my blog and I'm going t	22.1896
summary	What does this mean for regular folks like you and me	18.62
summary	It does this by reading your web site or blog and making connections between the social profiles you have	18.2541
summary	While Google is providing the API nobody is dependent on them for creating or storing our relationships	17.4752
summary	The social relationships that the API exposes are encoded in regular old HTML using the XFN and FOAF formats	16.3498

We perform the experiment on 20 bookmark pages which includes average 18 comments each. The average precision at 20 is 0.54 and the average recall is 0.19. (at a given cu-off rank, the recall value does not really matter.)

6. DISCUSSION

For the example above, the precision at 10 is the value of 0.9 and recall is 0.45. (When we retrieve more comments to 20 sentences, the precision is 0.7 and recall is 0.7.) This result only makes sense on the assumption that there are enough (over a half) sentences for summarization in the comment list. We call this 'summarization feature' in the social services like social bookmark service. In out experiment, there are over 51% summary sentences. If there are few related information, this approach may perform poor.

Besides, among the low ranked sentences, though we can find a good and more concise one, the algorithm cannot pick that up. The sentence "Google Social Graph API" is weighted 1.6259. We think this one is one of the best summaries. There are tendency in our sentence constructing technique that the longer sentence gets the more weight from extracted word.

On the other hand, in the example of blog titled "What is Web 2.0 – Tim Oreilly", the good summary is made of common word such as 'Web', '2.0' and so on. But to find out this good summary: "This article is an attempt to clarify just what we mean by Web 2.0" , considering 'Web 2.0' rather than 'Web' and '2.0' is one of the effective ways. This will be resolved by introducing word dictionary or the grammatical solution such as linguistic tree constructing.

In this paper, we have described an approach to blog summarization, which we call social summarization that uses the comment list in the social bookmark service such as del.icio.us. In turn, we have presented the experiment about measuring performance for this approach. This result clearly highlights the potential benefits of this summarization technique, which is seen to produce socialized information from many users, closely related to a human generated gold-standard.

In our experiments, sentence scores are sometimes very high because several words are repeated. Future direction of this approach is the consideration of sentence evaluation factors such as length, word's position and so on. In addition word extraction algorithm should be refined for more comprehensive study on comment pattern.

7. CONCLUSION

This paper presents a blog summarization technique using comments appeared in the social bookmark service. Based on an analysis of the comment classification in the social bookmark services, we found that users have propensity to make a summary for their bookmark contents. Our proposed solution finds representative words in comments set by term frequency weighting, grammatical tagging, and stop-word filtering. Using summary generated from the representative words, we evaluated performance of our work.

8. REFERENCES

[1] Gilad Mishne and Natalie Glance. 2006. Leave a Reply: An Analysis of Weblog Comments. In Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem (The Edinburgh, The Scotland, April 23-26, 2006)

[2] Meishan Hu, Aixin Sun and Ee-Peng Lim. 2007. Comments-Oriented blog Summarization by Sentence Extraction. In Proceedings of the 16th Conference on Information and Knowledge Management (Lisboa, Portugal, November 6-9, 2007)

[3] Del.icio.us. http://del.icio.us. 2008.02.04

[4] Digg. http://www.digg.com. 2008.02.04

[5] Christopher D. manning and Hinrich Schutze. 1999. Foundations of Statistical Natural Language Processing, The MIT Press.

[6] The Stanford POS Tagger. http://nlp.standford.edu/software/tagger.shtml. 2008.02.06

[7] A list of English stop words. http://en.wikipedia.org/wiki/Stop_words. 2008.02.06