FINDING SUITABLE KEYWORDS FOR A WEB PAGE FROM CACHES BASED ON SIMILARITY AND FREQUENCY

Yasuhiro Tajima and Yoshiyuki Kotani
Department of Computer and Communication Sciences, Tokyo University of Agriculture and Technology
Naka-chou 2-24-16, Koganei, Tokyo, Japan

Keywords: Metadata, Bayesian inference, Keyword generation.

Abstract: Meta data are most important entry in a web page for summarization, indexing, and so on. Unfortunately, there are many kind of matadata item but there are few guidelines for construct the metadata for a web page. We propose an metadata finding method for a web page by searching the internet caches and selecting suitable items for the target page. Our method is based on a bayesian method which is used in the area of text retrieval. We evaluate this method by an experiment to find a set of suitable keywords for a source web page. Compareing the original metatagged keywords and the system output, we obtain 74% precision and 76% recall. We can conclude that this method finds the tendency of metadata which is annotated to the pages similar to the target page.

1 INTRODUCTION

Metadata for a web page is important information and which is used for summarization, indexing or information retrieval. There have been suggested some guidelines for writing metadata, such as Dublin Core. These guidelines usually rule the entry types of metadata but does not define the content of data and description manner. In addition, Search Engine Optimization (SEO for short) makes confusion for smart use of metadata. Thus it tends to be a hard work that generating useful metadata for a human.

Automatic generation of metadata is one of the most expected study in the area of semantic web and some studies are exist for metadata retrieval(Jane Greenberg, 2005)(Jihie Kim, 2006). In (Jürgen Belizki, 2006), a metadata assignment method which is based on usage is proposed. In (Heiner Stuckenschmidt, 2001), a definition of a structure of metadata and its generation process is shown. There are also an application of such automatic metadata generation system(Paynter, 2005) and a specialized method for the area of image processing(Solomon Atnafu, 2002). However, these studies are usually based on an extraction method which takes metadata from the contents of the target file. Now, a set of keywords for a web page is significant for indexing and summarizing. But, it is also difficult because keywords may not be included in the target web page, thus we can not extract keywords only from one page data. For keywords metadata, we must find the suitable one with some background knowledge, because a set of keywords should contain an abstract word for the subject of the target web page and also contain a characteristic word to show a difference from other pages. It should need some inference mechanism from the background knowledge to make a suitable keywords for a web page. Thus, finding suitable keywords for a web page is also an important problem for semantic web.

In this paper, we show a metadata (especially for keywords) finding method for a web page by selecting suitable item from cached pages. In our method, the metadata added in the page which is similar to the target page will be selected. For an item of metadata, the probability whether the item is selected or not is decided by Bayesian method.

We evaluate this method by an experiment to find a set of suitable keywords for a source web page. Comparing the original meta tagged keywords and the system output, we obtain 74% precision and 76% recall. We can conclude that this method finds the ten-

Tajima Y. and Kotani Y. (2007).
FINDING SUITABLE KEYWORDS FOR A WEB PAGE FROM CACHES BASED ON SIMILARITY AND FREQUENCY.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 474-477
DOI: 10.5220/0001289204740477

dency of metadata which is added in the cache pages which are similar to the target page.

2 OUR METHOD

2.1 System Overview

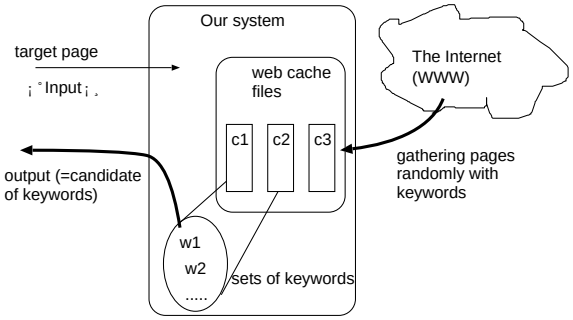

Figure 1: The overview of our system.

We have made a metadata finding system with our method. In Fig.1, we show the overview of our system. Our system consists of the following elements.

- s is the input target page file. This is the page which has no metadata and for which our system selects suitable metadata.
- C is a set of cache pages collected from WWW. This set C is contains only text data. Images and other format data will be omitted in our system.
- $c \in C$ is a page in the cache.
- K_c is the set of keywords annotated in the page c.
- W_c is the set of words which is contained in the page c.

With these items, metadata for the target page s is generated by the following steps.

1. Find similarity between the source page s and every cache page $c \in C$.
2. Select metadata of $c \in C$ based on the similarity.
3. Output the metadata for s.

In the following sections, we describe details of similarity calculation and selection algorithm.

2.2 Keyword Inference

Assume that W is a candidate set of keywords and let Z be the event that W is a set of keywords for s. Let $\bar{Z}$ be the event that W is not a set of keywords for s. Now, we consider the following value:

$$\frac{P(Z|W)}{P(\bar{Z}|W)}. \tag{1}$$

If we can find a set of words W which maximize (1), then W is suitable for a set of keywords for s. From Bayes theorem, we have

$$\log\left(\frac{P(Z|W)}{P(\bar{Z}|W)}\right) = \log\left(\frac{P(W|Z)}{P(W|\bar{Z})}\right) + \log\left(\frac{P(Z)}{P(\bar{Z})}\right).$$

Here, the second term $\log\left(\frac{P(Z)}{P(\bar{Z})}\right)$ has the same value for any W, thus we only consider the first term $\log\left(\frac{P(W|Z)}{P(W|\bar{Z})}\right)$.

Suppose that the occurrence probability of any word is identical each other. Then,

$$P(W|Z) = \prod_{u \in W} P(u|Z)$$

and

$$P(W|\bar{Z}) = \prod_{u \in W} P(u|\bar{Z}).$$

Now, the value $P(u|Z)$ can be approximated by the probability that u is used as a keyword in cache pages which are similar to s. In addition, $P(u|\bar{Z})$ can be approximated by the probability that u is not a keyword but it is contained in the page body of $c \in C$. Then, $P(u|Z)$ and $P(u|\bar{Z})$ can be written as

$$P(u|Z) = \frac{\sum_{c \in C, u \in K_c} \delta(s,c)}{\sum_{c \in C} \delta(s,c)} \tag{2}$$

and

$$P(u|\bar{Z}) = \frac{\sum_{c \in C, u \in W_c, u \notin K_c} \delta(s,c)}{\sum_{c \in C} \delta(s,c)}. \tag{3}$$

Here, $\delta(s,c)$ is the similarity between the target page s and a cache page $c \in C$. In this paper, we use the cos based similarity between s and c.

From these approximation, we can define the score $S(W)$ of a word set W as

$$\begin{aligned} S(W) &= \log\left(\frac{P(W|Z)}{P(W|\bar{Z}}\right) \\ &= \sum_{u \in W} \log\left(\frac{\sum_{c \in C, u \in K_c} \delta(s,c)}{\sum_{c \in C, u \in W_c, u \notin K_c} \delta(s,c)}\right) \end{aligned}$$

In Fig.2, we show an example of the above probability.

2.3 Yet Another Approximation

In equations (2) and (3), we have shown an example of approximation for $P(u|Z)$ and $P(u|\bar{Z})$. If $P(u|\bar{Z})$ is approximated by the probability that u is not a keyword but u is used in the page body among cache pages each of them contains u, then we can consider another approximation such that

$$P(u|\bar{Z}) = \frac{\sum_{c \in C, u \in W_c, u \notin K_c} \delta(s,c)}{\sum_{c \in C, u \in W_c} \delta(s,c)}. \tag{4}$$

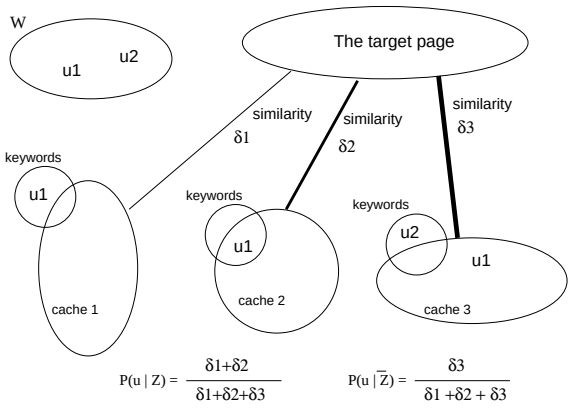

Figure 2: The overview of our system.

3 EXPERIMENT

We evaluate our method by an experiment as follows.

cache page number 1886 pages.

Every page in this set can be reachable from http://www.ipsj.or.jp by following the links. Also, every page in this set has metadata named "keywords". Thus, the crawler for our experiment searches all link paths from the page http://www.ipsj.or.jp, and takes keyword assigned pages into the cache set.

test page number 178 pages(set1) and 100 pages(set2).

From this set of pages, a target page is given to the system as input. The selection of these test set is random but we take care of variety of sites in the test set.

In Fig.3, we show the result of experiment of our method for the test data (set1). Here, n in these figures represents the size of words of keyword candidate. When $n = 1$, i.e. the output of our system contains just one word, then the precision is 1 and the recall is 0.1 for this test data. For the test data (set2), we have obtained that the precision is 0.8 and the recall is 0.08 with just one candidate.

The average number of keywords in the set of cache pages is 10.5 for the test data (set1). When $n = 10$, we can see that the precision is 0.78 and the recall is 0.72 for the test data (set1). On the other hand, for the test data (set2), we have obtained that the precision is 0.21 and the recall is 0.20 when $n = 10$.

In Fig.4, we show the result of experiment with the approximation in equation (4). From these graphs, we can read that if $n = 5$, i.e. the output of our system contains five words, then precision is 0.5 and recall is 0.3 for test data (set1). For the test data (set2), we have obtained that the precision is 0.2 and the recall is 0.1 when $n = 5$.

In both data set, (set1) or (set2), we can read that the approximation in the equation (4) is worse than another approximation.

4 CONCLUSIONS

We have shown a metadata finding method for a web page by selecting suitable item from web cache. In the evaluation, the result is 74% precision and 76% recall with data (set1) and $n = 10$.

For future works, inference of other metadata item, "subject" for example, is an interesting problem.

REFERENCES

Heiner Stuckenschmidt, F. v. H. (2001). Ontology-based metadata generation from semi-structured information. In *Proceedings of the First Conference on Knowledge Capture (K-CAP'01)*, pages 440–444.

Jane Greenberg, Kristina Spurgin, A. C. (2005). Final report for the amega (automatic metadata generation applications) project. In *University of North Carolina at Chapel Hill*.

Jihie Kim, Yolanda Gil, V. R. (2006). Semantic metadata generation for large scientific workflows. In *Proceedings of the 5th International Semantic Web Conference 2006 (ISWC2006)*, pages 357–370.

Jürgen Belizki, Stefania Costache, W. N. (2006). Application independent metadata generation. In *Proceedings of the 1st international workshop on Contextualized attention metadata: collecting, managing and exploiting of rich usage information(CAMA06)*, pages 33–36.

Paynter, G. W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 291–300.

Solomon Atnafu, Richard Chbeir, L. B. (2002). Efficient content-based and metadata retrieval in image database. In *Journal of Universal Computer Science*, volume 8, pages 613–622.

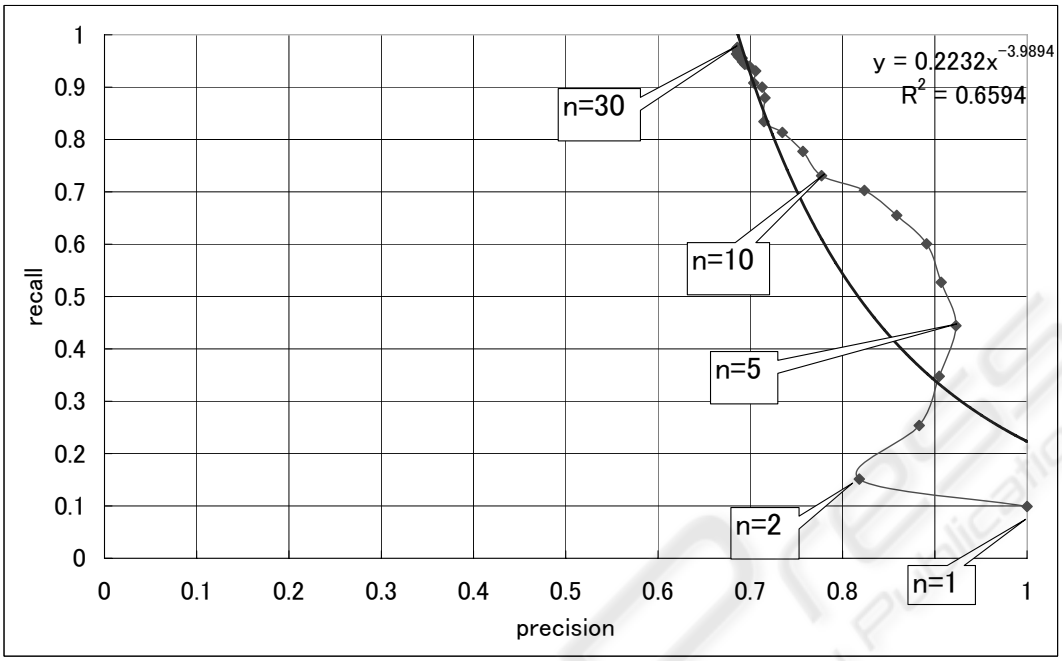

Figure 3: The result of experiment with (set1).

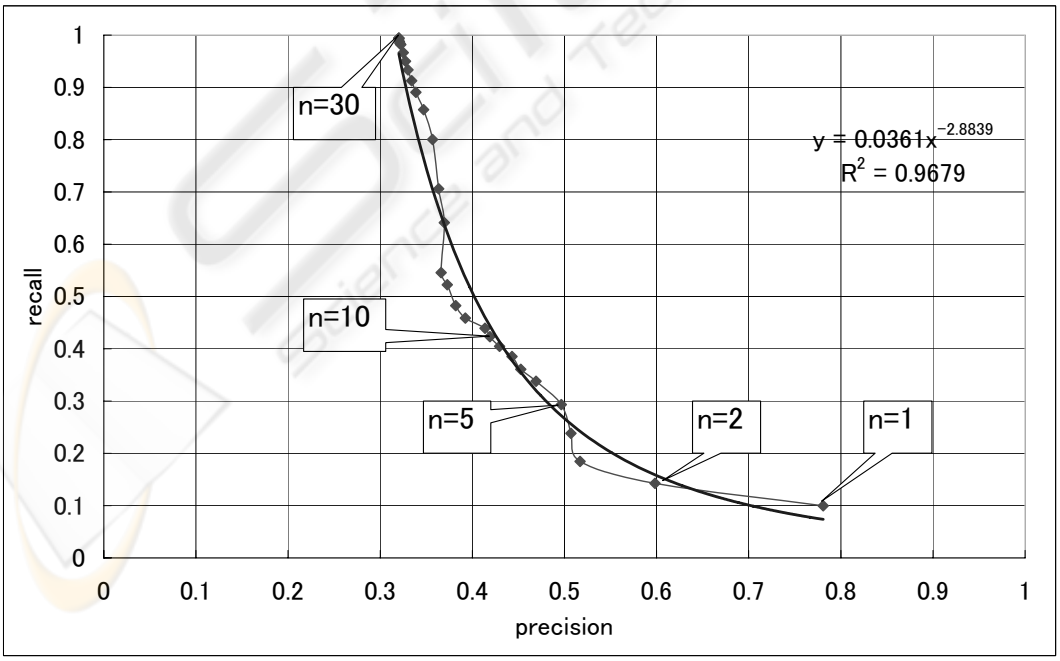

Figure 4: The result of (set1), another approximation.