Content Hole Search in Community-type Content

Akiyo Nadamoto
Konan University
Okamoto, Higashinada-ku,
Kobe, Japan
nadamoto@konan-u.ac.jp

Eiji Aramaki
The University of Tokyo
Hongou 7–3–1, Bunkyo-ku,
Tokyo, Japan
eiji.aramaki@gmail.com

Takeshi Abekawa
The University of Tokyo
Graduate School of Education
Hongou 7–3–1, Bunkyo-ku,
Tokyo, Japan
abekawa@p.u-tokyo.ac.jp

Yohei Murakami
National Institute of
Information and
Communications Technology
Hikaridai 3–5, Seika-cho,
Soraku-gun,Kyoto, Japan
yohei@nict.go.jp

ABSTRACT

In community-type content such as blogs and SNSs, we call the user's unawareness of information as a ”content hole” and the search for this information as a ”content hole search.” A content hole search differs from similarity searching and has a variety of types. In this paper, we propose different types of content holes and define each type. We also propose an analysis of dialogue related to community-type content and introduce content hole search by using Wikipedia as an example.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

General Terms

Algorithms

Keywords

Content Hole Search, Community, SNS, Blog

1. INTRODUCTION

Community-type content such as blogs and SNSs are representative examples of Web 2.0. In the community-type content, users sometimes enter heated discussions and their viewpoints become narrow. The community concentrates on one thing and loses information about their theme. In this case, we believe that taking notice of a lack of awareness is not only convenient for users but also for communities. On the other hand, today's web search techniques are primarily similarity searches that search for the information required by the user. When Web 2.0 became popular, users' viewpoints became increasingly narrow. Indeed, it is our contention that a new search technique that effectively increases users' content knowledge in relation to Web 2.0 is required. We need a next-generation search engine that will search for additional information regarding a topic that users might not know. We call this lack of information as ”content hole.” In this paper, we have attempted to extract and represent content holes from the discussion histories on SNSs and in blogs.

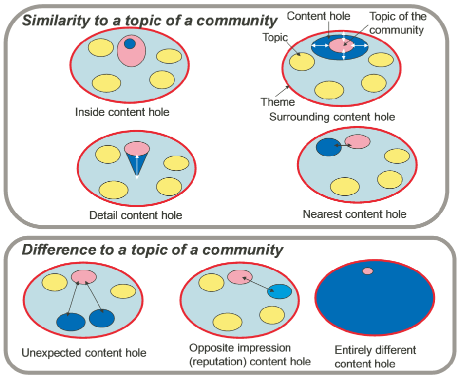


Figure 1: Types of Content Holes

A content hole search concentrates on differences, while ordinary information retrieval is generally based on similarities. We have compiled a number of images for the difference search, and we define 7 kinds of content holes[1]. In addition, all content hole searches use the same dialogue analysis logic. In this paper, we propose the dialogue analysis of community-type content; the first step in finding a content hole is to extract an ”unexpected content hole” by using Wikipedia.

2. CONCEPT OF CONTENT HOLE

The purpose of a content hole is to present information that is lacking within a given community. There are many kinds of content holes. Figure 1 shows an image of the content hole proposed in this paper. The community-type content consists of a Theme T and topic Sub_n $(n = 1, \ldots, n)$. T is a theme targeted by the community. Sub_n consists of a set of user comments C_i $(i = 1, \ldots, n)$ that have the same

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

topic as the theme. A theme consists of multiple topics, and a topic consists of one or more comments. We introduce our content hole by using the case of a topic consisting of a single comment in Figure 1. We have proposed seven content holes and have divided them into two kinds. One kind of content hole is similar to a community topic (Inside Content Holes, Surrounding Content Holes, Detail Content Holes, and Nearest Content Holes) and the other differs from a community topic (Unexpected Content Holes, Opposite Impression (Reputation) Content Holes, and Entirely Different Content Holes).

3. ANALYSIS OF THE DIALOGUE

Community-type content consists of a number of threads, with multiple dialogues within each thread. We have analyzed the threads and extracted a set of dialogues. We propose two types of relevant dialogues in a thread.

3.1 Content relevance

Content relevance refers to the similarity between two comments. For example, consider the following sentences: ”Which MP3 player is light and small?” and ”An IPod is extremely light and small.” We deduced that these sentences were probably related because the words ”*light*” and ”*small*” are present in both sentences. To calculate the content relevance, we use a co-occurrence-based similarity [2].

3.2 Function relevance

In the following case, there is little overlap of words between the two sentences; the relationship between these sentences comes in the form of a question and an answer. The following is an example of one such question and answer combination: ”Why does my iPod Nano occasionally stop?” and ”Because of the battery display.” To capture the functional relevance, we propose the following new measure: First, we built 3 databases using a set of comment pairs (Ps and Qs):

DB-A: a database of n-gram occurrences in Ps.

DB-B: a database of n-gram occurrences in Qs.

DB-C: a database of possible combinations of n : m-gram pair (possible n-grams in P: possible m-grams in Q) occurrences ($1 \leq n \leq 3, 1 \leq m \leq 3,$). For example, given the combination of P ”*How about an iPod*” and Q ”*Good idea,*” we obtained the n : m-grams.

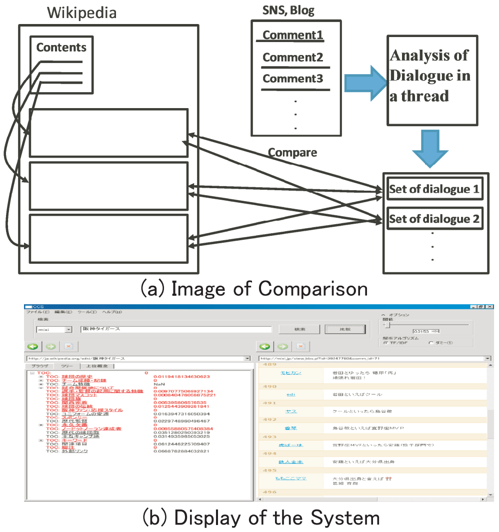
We define the functional relevance($REL_f(P,Q)$) using the following databases as follows:

$$REL_f(P,Q) = \sum_{p \in N_P} \max \sum_{q \in N_Q} CPMI(p,q), \qquad (1)$$

where N_P is a set of n-grams in P, N_Q is a set of n-grams in Q, and $CPMI$ is defined as follows:

$$CPMI(p,q) = \begin{cases} 0 & \text{if } H_c(p \cap q) \leq c, \\ \log \dfrac{\frac{H_c(p \cap q)}{M}}{\frac{H_a(p)}{M}\frac{H_b(q)}{M}} & \text{otherwise,} \end{cases} \qquad (2)$$

where $H_a(p)$ is the number of occurrences of n-gram p in DB-A, $H_b(q)$ is the number of occurrences of n-gram q in DB-B, and $H_c(p \cap q)$ is the number of occurrences of the n-gram pair p and q in DB-C. We filtered out queries that returned less than a threshold limit of c number of documents to avoid small-number noise. M is the number of comment pairs.



(a) Image of Comparison

(b) Display of the System

Figure 2: Prototype System

4. A CONTENT HOLE SEARCH

In this paper, the first step in finding a content hole is to extract an Unexpected Content Hole by using Wikipedia’s table of contents. Wikipedia is written by the general public and thus contains multiple perspectives as well as colloquialisms. We extracted the Unexpected Content Hole using Wikipedia because it contains content that is written from multiple perspectives. We used Wikipedia’s table of contents as a structure of Wikipedia. In particular, we compared each dialogue set in a thread with each passage in the table of contents, as shown in Figure 2(a). First, we calculated the morphological analysis and computed the weight of all nouns in each passage by using TF/IDF. We then computed the similarity between each dialogue set and each passage in the table of contents by using a cosine vector. The passage from Wikipedia’s table of contents is smaller than the threshold of the content hole. Figure 2(b) shows the interface of the prototype system. In this system, the user inputs the keyword, and the system searches for communities that discuss the keyword. The user then selects the community, and the system calculates and presents that community’s content hole in red characters in the left-hand window.

5. CONCLUSION

We proposed a content hole search for community-type content. We also proposed an analysis method for dialogue within the various threads in a community. Furthermore, we describe a manner in which a content hole search can be performed using Wikipedia.

6. REFERENCES

[1] A Nadamoto, E Aramaki,et. al., Yohei Murakami, ”Searching for Important but Neglected Content from Community-type-content”, SITIS, pp.161–168, 2008

[2] Danushka Bollegala, et. al., ”Measuring Semantic Similarity between Words Using Web Search Engines”, Proceedings of 16th International World Wide Web Conference (WWW 2007), pp.757-766, 2007