Serendipitous Search via Wikipedia: A Query Log Analysis

Tetsuya Sakai
Microsoft Research Asia
tetsuyasakai@acm.org

Kenichi Nogami
NewsWatch, Inc.
noga@newswatch.co.jp

This work was done while the first author was at NewsWatch.

ABSTRACT

We analyse the query log of a click-oriented Japanese search engine that utilises the link structures of Wikipedia for encouraging the user to change his information need and to perform repeated, serendipitous, exploratory search. Our results show that users tend to make transitions within the same query type: from person names to person names, from place names to place names, and so on.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors

Keywords

Search engine, clickthrough, users

1. INTRODUCTION

Exploratoy search [5], which is beginning to receive a lot of attention, covers not only querying and producing a ranked output, but also forming and changing information need. Twenty years ago, Bates [2] proposed the *evolving search* and *berrypicking* models for handling change in information need and what people now call exploratory search.

In September 2008, we released a click-oriented Japanese search engine called *KotobaNoUchu* (Galaxies of Words), which utilises the link structures of Wikipedia for encouraging the user to change his information need and to perform repeated, serendipitous, exploratory search. We thereby aim to provide the user with a lot of *useful* information that may not be *relevant* to his initial information need.

We analyse the query log of Galaxies of Words to see how users move from one query to the next on our clickable Galaxy interface, which visualises the link structures of Wikipedia. Studies on Japanese query logs include [1, 3], but none of them addresses the issues of substantial change in information need. In fact, to our knowledge, query log studies in general have tended to overlook these issues.

2. GALAXY: THE VISUALISED WIKIPEDIA

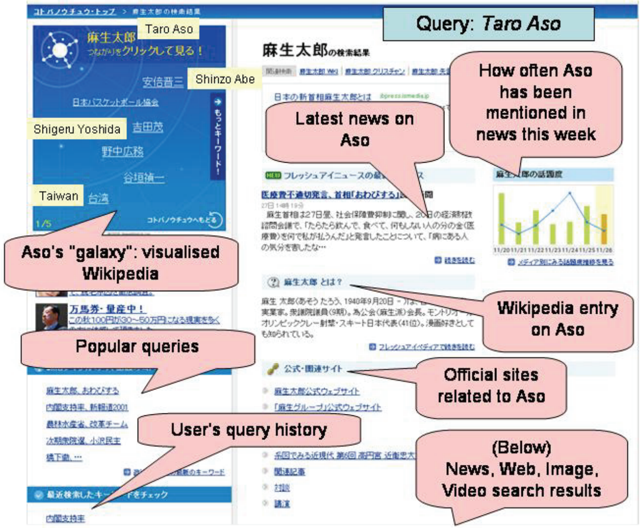


Figure 1: A sample search output of Galaxies of Words (`http://kotochu.fresheye.com/`).

Figure 1 shows a sample search output of Galaxies of Words. Here, the query is "Taro Aso", the current prime minister of Japan. The search output contains several ranked lists (news, Web, image, video search output etc.), and the Galaxy, which is the square area shown on the top left corner. In Taro Aso's Galaxy (or rather, solar system), the Wikipedia entry "Taro Aso" is the sun, and the other Wikipedia entries linked from the Aso entry are the planets. If the user puts his mouse over the "Shigeru Yoshida" planet, the Galaxy shows a pop-up window, which contains a context within the Aso entry that mentions Yoshida. Thus the user can discover that Yoshida, who was also once a prime minister, is Aso's grandfather.

If the user is interested in Yoshida, he may click on it in the Galaxy. Then, a new search output is produced, with "Shigeru Yoshida" as the query this time. Hence Yoshida becomes the sun in the new galaxy. In this way, Galaxies of Words stimulates the user's curiosity, and encourages change in the user's information need. If the user thus keeps clicking on different planets, he may find pieces of useful and/or interesting information during the process. From the viewpoint of our company, this means many page views and therefore more profit through advertising.

In general, a Wikipedia entry contains many links, so the Galaxy contains a "next page" button for showing more planets. In the current version of Galaxies of Words, we use some heuristics for prioritising the planets: We favour person names, organisation names and so on, and we also favour links that occur frequently within the Wikipedia page. In this study, we conduct a query log analysis to see what kind

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

of planets we should present to the user in order to make the serendipitous/exploratory search more successful.

3. QUERY LOG ANALYSIS

We analysed the query log for the entire month of October 2008 (one month after the site was released). We had 806,772 *records*, where each record is a quadruple of the form: *timestamp*, *Cookie*, *query before transition*, *query after transition*. We obtained 462,891 user *sessions* (409,815 users) using thirty minutes as the timeout threshold as in previous studies. Among the above data, only 20,311 sessions contained a transition to a planet within the Galaxy. The transition length was 1.19 on average, and 80 at maximum. (This user clicked on a planet 80 times in a row!)

We obtained 13,258 unique queries from the above 20,311 sessions, and manually classified them into the following categories: **PERSON**, **ORG** (organisations), **TITLE** (books, movies, etc.), **GROUP** (baseball teams, pop groups, etc.), **PLACE**, **ADULT**, **EVENT**, **PROFESSION**, **INCIDENT**, **DISEASE**, **PRODUCT**, **WEBSITE**, **LAW**, **ANIMAL**, **TIME** and **NO CATEG** (no category). We then analysed how users move from one query type to another on the Galaxy interface. Figures 2-6 show some selected results.

Although the distribution is biased towards **PERSON** planets and **ORG** planets because they have the highest chance of being shown to the user, it can be observed that users tend to move within the same query type: from person names to person names, from place names to place names, and so on. This is true for other query types not shown in the figure: For example, the chance of moving from an **EVENT** query to another **EVENT** query is about 24%; that of moving from a **DISEASE** query to another **DISEASE** query is about 38%.

4. CONCLUSIONS

We showed that, on our clickable Galaxy interface for serendipitous search, users tend to make transitions within the same query type: from person names to person names, from place names to place names, and so on. We plan to utilise this finding for selecting good planets and for pruning planets that are probably less useful. We also plan to customise the Galaxy based on clicks for each user.

5. REFERENCES

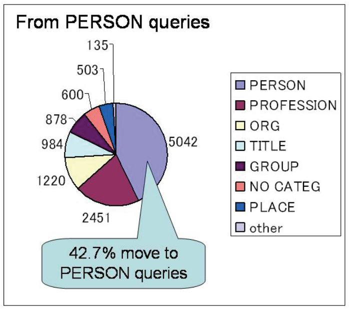
[1] Baeza-Yates, R., Dupret, G. and Velasco, J.: A Study of Mobile Search Queries in Japan, *WWW 2007 Workshop on Query Log Analysis: Social and Technological Challenges* (2007).

[2] Bates, M. J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface, *Online Review*, 13(5), pp. 407-424 (1989).

[3] Jones, R., Bartz, K., Subasic, P. and Rey, B.: Automatically Generating Related Queries in Japanese, *Language Resources and Evaluation*, Vol.40, No.3-4 (2006).

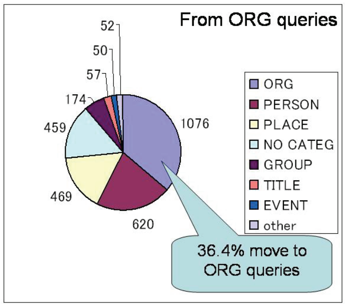
[4] Sakai, T. *et al.*: Design and Development of an Exploratory Search System based on Clickthroughs (in Japanese), *Forum on Information Technology 2008*, pp.1-4 (2008).

[5] White, R. W. *et al.* (eds.): *Proceedings of the ACM SIGCHI 2007 Workshop on Exploratory Search and HCI: Designing and Evaluating Interfaces to Support Exploratory Search Interaction* (2007).



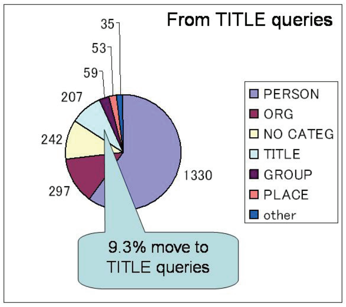
42.7% move to PERSON queries

Figure 2: Transitions from PERSON queries.



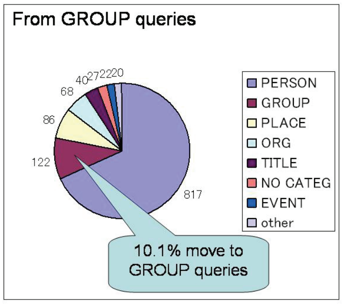
36.4% move to ORG queries

Figure 3: Transitions from ORG queries.



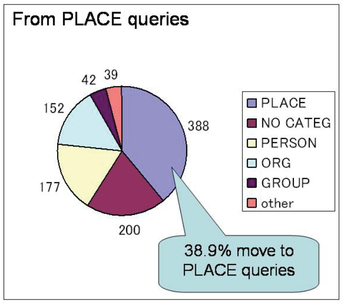
9.3% move to TITLE queries

Figure 4: Transitions from TITLE queries.



10.1% move to GROUP queries

Figure 5: Transitions from GROUP queries.



38.9% move to PLACE queries

Figure 6: Transitions from PLACE queries.