A General Classification of (Search) Queries and Terms

Nadine Schmidt-Maenz and Martina Koch
Institute for Decision Theory and Management Science, University of Karlsruhe (TH), Germany

Abstract

Web information systems are the most popular service for online users to retrieve data (pages, images, or files) on the internet. In order to get a more profound insight into the method and the target of web searches, we have tracked the queries that have been entered at the Lycos search engine over several months. The analysis of this vast amount of empirical data provides us with a deeper understanding of online users' behavior.

In this paper, we focus on time-dependency in the usage of terms. Furthermore, we use aspects from the human online information processing in search engines and from topic detection in documents to discuss a general classification of terms. As a result, we find time-dependent clusters of (search) terms around particular subjects. Based on these findings, strategies for the design of search engines and web pages which focus on the (information) consumer will be developed. The basic classification which is presented in this paper can also be leveraged as an indication of the adequacy of mathematical models for simulating the usage of terms over time.

1. Introduction

As a well-known subgroup of web information systems, search engines play a very important role for discovering new web pages or content on the web. In several studies [1, 3, 4, 8, 10] researchers examined diverse facets of online searching behavior of web users and summarized the main facts, regarding vocabulary growth or the average length of (search) queries. These are complemented by studies about the most frequently used search terms [9, 11]. For a more detailed literature overview see [6] and [10].

Having analyzed a huge amount of search queries, we found out, as one of our results, that the ranked listing of the most frequently used (search) queries has not changed significantly over time. For that reason we frame following research questions: 1. Is there a better way to identify the most popular terms? 2. Does a basic classification of terms exist dependent on their appearance in the data set? 3. What are the general characteristics of the usage of terms? 4. Is it possible to extract topic clusters of (search) queries and terms based on these findings?

In this paper, we introduce a procedure for discovering the distribution of (search) terms, which is sensitive to periodical changes as well as to the interest of web users in special search topics. This procedure facilitates the classification of search terms in general time-dependent clusters. In the next step we take our findings on the information processing of web users as an input for a classification of search topics and give examples of these clusters with their enclosed words.

In the following, we first introduce basic definitions and descriptions. By highlighting some key facts about (search) queries and terms, we give a short overview of the data which we used. The next section is dedicated to examples of the top (search) queries and terms and, in section 4, we discuss the distribution of (search) terms over a period of time. A summary of our findings on human online information processing to deduce a general classification of (search) terms based on the frequency of their appearance during the observation period is located in section 5. In the section afterwards we introduce decision rules for identifying the topic and member items of clusters and give some examples of clusters which were determined by these rules. The final section provides a summary of our findings, conclusions, and an outlook for further research.

2. Basic Definitions and Descriptions

General search engines such as Lycos maintain their own index of data available on the web (web pages). This index is automatically created by web crawlers. In addition to general search engines, there are other web information systems such as meta search engines (Metager) or directories (Open Directory Project) [6]. Meta search engines do not create their own index. Instead, they access the indices of a range of other search engines simultaneously. The database of directories is generated manually. Webmasters need to submit an application for the incorporation of their web pages, which are reviewed before they are added to the directory. Some search engines have a 'live ticker', 'live search', or alternatively a 'spy function' enabling users to see the current (search) queries of other users. A (search) term is an uninterrupted sequence of characters (letters or

Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)
0-7695-2497-4/06 $20.00 © 2006 IEEE

numbers). Operators (OR, AND or NOT) are logic constructs to combine terms and are themselves not treated as such. Several terms enfolded by quotes, also a sort of operators, form a phrase which is treated as a single term by search engines. A non-empty (search) query represents one or more terms a person has typed in a search engine's search interface with operators or spaces in between. A (search) topic is a label which combines different (search) queries and the terms embodied of the same subject. In the following section we introduce some mathematical explanations for our data set.

The observation period is denoted by [1;T] and t describes one equidistant interval in this period (hours, days, or weeks). V stands for *vocabulary* (terms ever used) and S for *(search) queries* (ever used). $\vec{V}$ is the N-dimensional vector where every dimension n=1, …, N represents a term that has appeared at least once in the queries observed. The values v_n stored in each dimension are the number of occurrences of the term n until the end of the observation period. The sum over n=1, …, N of v_n is referred to as the gross number of term occurrences (Γ_v) and N is the net number of terms at the point of time T. The average occurrence $\bar{v}$ is calculated by the quotient of the gross and net number. $\vec{S}$ is the M-dimensional vector of (search) queries and s_m is the number of occurrences of a (search) query m=1, …, M. The gross (Γ_s), the net (M), and the average ($\bar{s}$) number of (search) query occurrences are defined equivalently to the measures of term occurrences mentioned above.

We observed the 'live search' of Lycos Europe for a period of eight month from 08/14/2004 to 04/14/2005 by a program which automatically stored the queries shown on that web page. Here, we only present the key data of our data set. A more detailed description of how to observe search engines' live tickers is illustrated in [6]. One time interval t equals one day. Our observation period is 244 days long such that $1 \le t \le 244$. The gross number of queries Γ_s in this data set is 118,729,604; M equals 19,598,122 and the average occurrence of (search) queries $\bar{s}$ is about 6.1 times. Regarding the terms in these queries observed, Γ_v is 205,937,248; N is 7,806,448, and the average term occurrence $\bar{v}$ is about 26.4 times. The average appearance of (search) terms is much higher than that of (search) queries. Thus, it is more probable that a term reoccurs a second time. In this data set 9.99% of all queries appeared only once (5.33 twice and 3.46 three times). With respect to the net number of queries, 60.52% occurred only once. Hence, there are only a few queries with a very large number of occurrences. The average length of queries is 1.7 which is similarly reported in other web user studies [8, 10]. The usage of operators in total is very small (< 3%), quotes are the mostly used operator to refine queries with a fraction of 1.7%.

3. Top (Search) Queries and Terms

In this section we give examples of the top (search) queries and terms after two different time frames of the observation period. Table 1 shows the top 10 queries after three months in November 2005 and then after six more months in February 2005. It is evident that the top queries consist of only one term and that they are not very helpful to find information in the web, since they contain only one very general term. Furthermore there are only slight changes with respect to the rank of queries, and no real shifts are noticeable. All in all, the top 10 list remains stable over a period of several month.

Table 1. Top-10 of (search) queries

Rank	November 2005	February 2005
01	Lycos	Lycos
02	Link:http://www.	Sex
03	Sex	Link:http://www.
04	Hentai	Hentai
05	Porno	Porno
06	Ebay	Ebay
07	Google	Erotik
08	Erotik	Google
09	Plexiglasgehaeuse	Fkk
10	Christina Aguilera	Christina Aguilera

Table 2 shows the top 10 of terms in queries after three months in November 2005 and after six months in February 2005. It is obvious that some top terms correspond to the top (search) queries. In addition there appear many fillers like *in*, *der*, *für*, or *und*.

Table 2. Top-10 of (search) terms

Rank	November 2005	February 2005
01	Lycos	Lycos
02	Sex	Sex
03	In	In
04	Der	Der
05	Link:http://www.	Und
06	Und	Hotel
07	Hentai	Von
08	Porno	Für
09	Hotel	Porno
10	Für	Free

But both, top 10 lists of queries and terms are not recommendable to conduct marketing campaigns or to adapt optimized caching strategies. However, it is interesting that people often use fillers to formulate

(search) queries such as 'Hotel in Berlin'. The conclusion is, that people in bulk do not understand the functioning of search engines. Fillers do not narrow a query, because search engines usually ignore them.

Since top 10 lists aren't informative to detect interesting topics, we will have a look at the distribution of occurrences of terms during our observation period. The results will be presented in the next section. From now on, we only consider (search) terms exclusively, since (search) queries on average are very short and since it is more likely that a term might appear again.

4. Analyzing (Search) Terms Over Time

To learn about the distribution of terms over an observation period it is helpful to visualize the amount of terms which appear in exactly one, two, three, or more time intervals. In dependence on the procedure reported in [15], we introduce a more detailed notation and visualize the distribution of our data set. With the definitions mentioned in the second section, we define characteristic parameters for the distribution of terms as follows:

- v_{nt} is the quantity of term n in time interval t,
- f is a fixed threshold depending on the search engine observed,
- $a_{nt}^f = \begin{cases} 1, & v_{nt} \geq f \\ 0, & \text{otherwise} \end{cases}$ is the indicator whether term n was frequent in time interval t,
- $c_n^f = \sum_{t=1}^{\mathrm{T}} a_{nt}^f$ is the number of time intervals in which a term n was frequent,
- $e_n^f(x) = \begin{cases} 1, & c_n^f = x \\ 0, & \text{otherwise} \end{cases}$ $x = 1, ..., \mathrm{T}$, is the indicator that term n was in exactly x intervals frequent,
- $p^f(x) = \sum_{n=1}^{\mathrm{N}} e_n^f(x)$ is the number of terms which were in exact 1,2, …, T time intervals frequent.

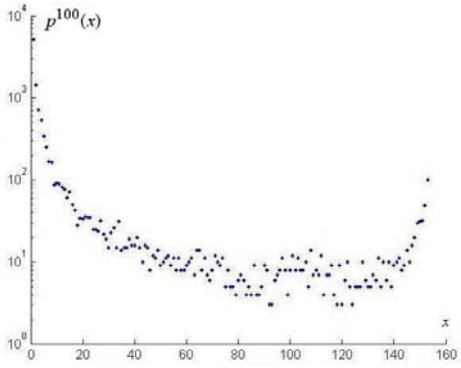

Figure 1. Function $p^{100}(x)$ of our data set

Figure 1 gives an example for the distribution of terms with T=160 , f=100 and one time interval t is equivalent to one day. We here choose a smaller time frame (T=160) to achieve a readily identifiable figure.

The characteristic of the discrete distribution of terms over a time period is that we have many terms which only appear in one or two time intervals and again many which appear in (nearly) every time interval.

Those terms which appear in nearly every time interval t, here, i.e. days, are defined as ***evergreens*****,** while those which appear only in one interval are called ***mayflies***. With this perception in mind we define three main time dependent clusters, i.e., the set of terms of mayflies (M) and the set of evergreens (E), with

- $\delta_1, \delta_2 \in [0;1] \wedge \delta_2 >> \delta_1$,
- $\mathsf{M} = \left\{ n \mid c_n^f \in [1; \delta_1 * \mathrm{T}] \right\}$, and
- $\mathsf{E} = \left\{ n \mid c_n^f \in [\delta_2 * \mathrm{T}; \mathrm{T}] \right\}$.

The set of terms in between is referred to as the ***midfield***. The cluster of mayflies is not of interest for caching strategies. Here, it is advisable to have a look at the set of evergreens. For a threshold of f=400 and $\delta_2 = 0.9$, we obtain as a set of evergreens

- $\mathsf{E} = \left\{ n \mid c_n^{400} \in [219.6; 244] \right\}$.

This set has 61 elements. Manually, we subdivide this cluster in different topics. The largest subcluster is labeled 'erotic' and contains 20 elements.

The second largest subcluster with 17 elements embodies wrongly used operators and fillers.

The third noteworthy subcluster contains 10 terms dealing with 'multimedia' topics like *software*, *online*, *chat*, *music,* or *mp3*.

The other subclusters deal with topics labeled 'shopping' (5), 'travel' (4), names of 'search engines' (3), and 'other' (2). It is remarkable that with the knowledge of these evergreens in (search) queries it is possible to implement optimized caching strategies.

For a larger threshold of f=1,000 and the same δ_2 as above the set of evergreens is almost the same as the top 10 lists mentioned in section 3.

We now have three time dependent clusters of terms with a huge midfield where interesting topics might be discoverable. To detect topics in the midfield, we have to examine how people process information online.

5. Human Online Information Processing

In this section, we elaborate profound aspects about human online information processing. We will give examples of how people search online for information of different importance and actuality.

Figures 2, 3, and 4 show the quantity of different strings in (search) queries per day (y-axes: v_{nt}) during

our observation period (x-axes: the accurate date), likewise patterns and curves are illustrated in [12].

Figure 2 shows two curves with high peaks from time to time ('kino' = cinema, 'kinoprogramm' = cinema program). The light gray curve oscillates about a higher level than the dark gray curve. The latter shows the same peaks but a very low curve progression in between. This pattern is caused by the fact that new movies appear periodically.

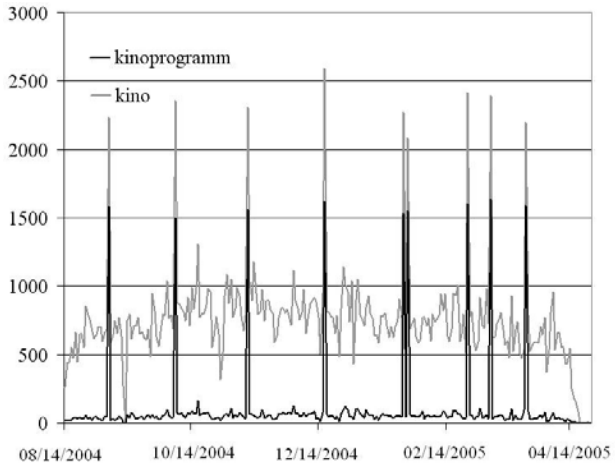

Figure 2. Characteristic line of events

Both lines can be seen as signals of human information interest. A characteristic information signal line for evergreens is generally of a high constant value with little variability.

It serves as an upper bound of random noise (s. figure 3: 'reise' = travel and 'flug' = flight). Curves with the same characteristics but settled on a very low level, close to zero, illustrate a lower bound of random noise.

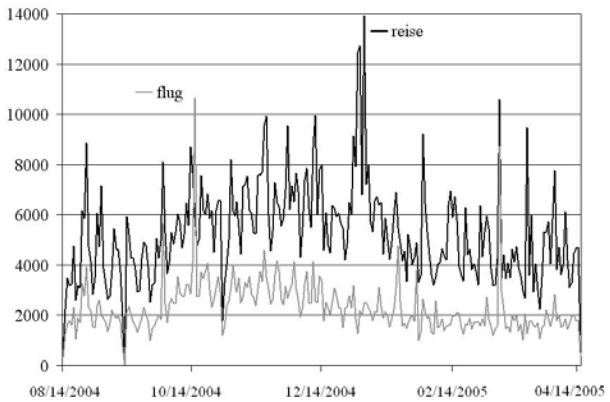

Figure 3. Characteristic line of evergreens

Accordingly, the threshold f is a parameter to cut off these curves of lower random noise. The area in between the upper and lower random noise form the ***information bandwidth***.

Some signals burst out of this information bandwidth once in a while as shown in figures 2, 3, and 4. This happens if terms are of a particular interest, or importance. These peaks occur if, on the one hand, a recurrent ***event*** (a baseball game, Olympic Games, or elections) is imminent or, on the other hand, if an unpredictable ***impulse*** occurs (an earthquake, a tsunami, or other natural disasters).

In [5] an event is defined as an unique occurrence that happens at a specific time. To that definition, we add impulses to differentiate between expected and unexpected events.

Impulses and events have one main different attribute: Impulses have a fixed starting point since an earthquake occurs suddenly. Impulses are of interest from that starting point on and disappear again with cascading information interest.

Then again, Events have a fixed endpoint like the day of election or the last day of Olympic Games.

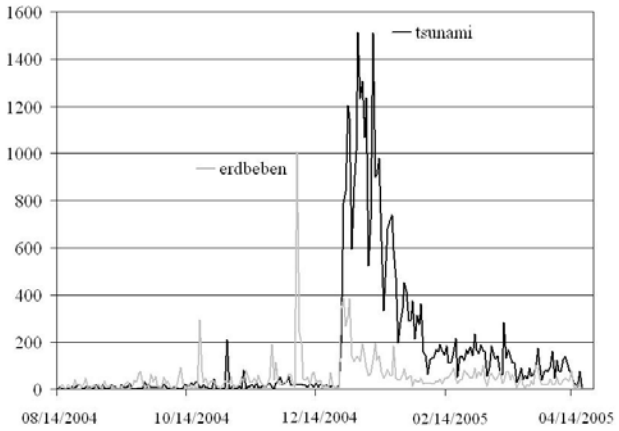

Figure 4. Characteristic line of impulses

Figure 4 shows the impulses ('erdbeben' = earthquake, tsunami). This first peak of 'erdbeben' occurred one day after a relatively hard earthquake in Germany and was of importance for only one day. The tsunami after a very severe seaquake in the Indian Ocean 2004 had a much longer information signal, where 'erdbeben' was again of interest.

As mentioned above, events reoccur in a periodical manner which is shown in figure 2. If one aggregates the number of occurrences of impulses and events we obtain values of c_n^f which are located in the midfield.

Some mayflies become impulses, when sudden reoccurrence is possible. There is a smooth transition in between.

Furthermore, an impulse can become an event, if it reoccurs in a periodical manner with a fixed deadline or if advance notices were published, e.g., if a hurricane would be expected.

The next task will be to estimate those transient probabilities and characteristic functions for evergreens. And to learn about the impact of advance notices.

6. Detecting Topics of (Search) Terms

The basic idea of topic detection and tracking is to identify new or similar stories which are arriving continuously in a database. Another task is to classify the following incoming documents in an online process. Further, there are approaches to visualize such a data stream concerning a special topic. The detection of new topics is also one principal task and is called first story detection. An introduction to topic detection and tracking (TDT) in documents and a literature review is given in [5]. In [14] three main properties are specified to classify associated news in data streams:

- Associated news appear nearly at the same time and contain the same vocabulary.
- A temporal gap between explosively increasing quantities of news, each with same content, indicates a new event.
- New topics in documents often become evident through new vocabulary.

The circumstances in the area of (search) queries are different, since they are very short: they contain only a few keywords. But it is possible to transfer some aspects to find topics of (search) terms. These aspects are the temporal proximity of occurrences of terms dealing with the same topic and new terms in the vocabulary. Further we can use the findings of human information processing to detect topic clusters around impulses and events after we have determined the evergreens. For this we introduce three new variables which define the first and last appearance of a term during an observation period and the intensity which expresses the percentage of intervals a term appeared in between the first and the last occurrence:

- $\text{first}_n^f = \min\left\{t \mid a_{nt}^f \geq f\right\}$,
- $\text{last}_n^f = \max\left\{t \mid a_{nt}^f \geq f\right\}$, and
- $in_n^f = \dfrac{c_n^f}{\text{last}_n^f - \text{first}_n^f + 1}$.

To detect topic clusters of particular impulses we formulate the following decision rule with f=400:

- first_n^{400} := first time interval in which the impulse occurred **AND**
- $\text{last}_n^{400} > \text{first}_n^{400}$ **AND**
- $in_n^{400} > 0.5$

We choose first_n^f dissimilar from last_n^f to disregard mayflies which occurred only in one time interval t. In our observation period a few known impulses and events occurred, such as the *tsunami* in the Indian Ocean 2004 (t=137) and the death of the *pope* John Paul II (t=232). With the decision rule described above we try to gain two keyword clusters: the *tsunami* and the *pope* cluster. The *tsunami* cluster also contains *dolpo* and *phuket*. While the *pope* cluster only contains *Johannes* in addition. These clusters are very small and do not contain any other terms. To discover term clusters of events, we design another decision rule based on our findings about human information processing:

- last_n^{400} := time interval of the event's target date **AND**
- $\text{first}_n^{400} < \text{last}_n^{400}$ **AND**
- $in_n^{400} > 0.5$

We here recognized four events: *Christmas*, *Halloween*, the *Olympic Games* 2004, and *Valentine's Day*. The *Christmas* cluster contains also:

- Weihnacht (Christmas)
- Weihnachtsbilder (Christmas pictures)
- Weihnachtsgedichte (Christmas poems)
- Weihnachtskarten (Christmas cards)
- Weihnachtsmann (Santa Claus)

The other clusters do not contain any term in addition to the one which indicates the event itself.

7. Summary and Conclusions

After having given some basic definitions, we demonstrated that top ten lists remain stable over a period of several months. We therefore introduced a fundamental time-dependent categorization of terms in mayflies and evergreens. In the next step, we defined the information bandwidth and signals of human information interest. With these findings about human online information processing we gave a definition of impulses and events.

Remarkable is that only particularly severe impulses form term clusters at all. Common impulses and events do not have the power to send information signals strong enough to form keyword clusters.

Noteworthy is also that only a few terms are associated with impulses or events. We showed that very simple decision rules are sufficient to extract small term clusters around a particular topic concerning an event or impulse. The variables and the decision rules defined have to be extended to detect weekly or monthly impulses and events. Furthermore, it is worthwhile to develop more

comprehensive procedures to generate more precise topic clusters.

The best marketing strategy for gaining online visibility is the hypothetical potential to induce an impulse concerning a product or a web page [7]. The flash of Janet Jackson resulted in such a burst of queries [2].

However, it is not always possible to induce such impulses. Instead, it is recommendable to pay attention to potential impulses in order to win the overwhelming interest of online searchers with specialized content or keywords on the web page, or rather to rent Adwords regarding an impulse [7]. A news area which links to web pages with news related content should be offered by search engines as well as portals. Building up a news index which is updated in very short time intervals is another recommendable strategy.

Leveraging the identification of evergreens, caching strategies can be developed as presented in [13]. Search engines can build an exclusive index which is only based on pages that deal with evergreen (search) topics and another which deals with non-evergreen topics. Such a divided index increases the capacity for new queries and decreases the response time, which is of particular importance when an impulse occurs. This insight is of high relevance for the design of online portals as well, since two main evergreens are erotic and travel subjects. It is also important to know evergreen topics for marketing campaigns or for generating increased traffic on pages. Another possibility is not to operate only with those common terms but to combine common evergreen terms with exotic ones.

All in all, we presented a very simple but effective procedure to examine those short (search) queries and to detect some topic clusters of (search) terms. Beyond this, our results reveal some fundamental aspects about human online information processing which will help to improve online services. It is of particular interest to simplify the online search, since it is not generally known, how search engines process (search) queries.

Finally, the most important outcome of our results is that evergreens, events or impulses should be described with different mathematical models. For evergreens, a broad time series analysis can be conducted, while events or impulses should be described by other models such as Fourier analysis [12]. But there is definitely no mathematical model which adequately describes the functionality of all terms in only one model. Our next step will be the further formalization of our findings.

8. References

[1] Beitzel. St., Jensen, E., Chowdhury, A., and Grossman, D. (2004): Hourly Analysis of a Very Large Topically Categorized Web Query Log, Proceedings of the 2004 ACM Conference on Research and Development in Information Retrieval (ACM-SIGIR), Sheffield, UK, July 2004.

[2] Charny, B. (2004): Janet Jackson's Flash Dance Tops Web Search, published on ZDNet News: February, 2004, http://news.zdnet.com/2100-2513_22-5153330.html (last visited 09/15/2005).

[3] Hoelscher, Ch. and Strube, G. (1999): Searching on the Web: Two Types of Expertise, Poster Proceedings of SIGIR'99, 305-306.

[4] Jansen, B., Spink, A., and Saracevic, T. (2000): Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web, Information Processing and Management, 36(2), 2007-2027.

[5] Kleinberg, J. (2006): Temporal Dynamics of On-Line Information Streams, to appear in: Garofalakis. M., Gehrke, J. and Rastogi ,R. (2006): Data Stream Management: Processing High-Speed Data Streams, Springer.

[6] Schmidt-Maenz, N. and Koch, M. (2005): Patterns in Search Queries, in: Baier, D., Decker, R., and Schmidt-Thieme, L. (Eds.) (2005): Data Analysis and Decision Support. Springer, Heidelberg, 122-129.

[7] Schmidt-Maenz, N. and Gaul W. (2005): Web Mining and Online Visibility, in: Weihs, C.; Gaul, W. (Eds.) (2005): Classification - the Ubiquitous Challenge. Springer, Heidelberg, 418-425.

[8] Silverstein, C. and Henzinger, M. (1999): Analysis of a Very Large Web Search Engine Query Log, ACM SIGIR Forum, 33(1), 6-12.

[9] Spink, A. and Gunar, O. (2001): E-Commerce Web Queries: Excite and Ask Jeeves Study, First Monday, Peer-Reviewed Journal on the Internet, 6(7).

[10] Spink, A. and Jansen, B. (2004): Web Search: Public Searching of the Web, Kluwer Academic Publishers.

[11] Spink, A., Jansen, B., Wolfram, D., and Saracevic, T. 2002. From E-Sex to E-Commerce: Web Search Changes. IEEE Computer. 35(3), 107 – 111.

[12] Vlachos, M., Meek, Ch., and Vagena, Z. (2004): Identifying Similarities, Periodicities and Bursts for Online Search Queries, Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, Paris, France, ACM Press, 131-142

[13] Xie, Y. and O'Hallaron, D. (2002): Locality in Search Engine Queries and Its Implications for Caching, Infocom (2002), http://www-2.cs.cmu.edu/~ylxie/papers/infocom02.ps.

[14] Yang, Y., Pierce, T., and Carbonell, J. (1998): A Study on Retrospective and On-line Event Detection, Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, 28-36.

[15] Zien, J., Meyer, J., Tomlin, J. and Liu, J. (2000): Web Query Characteristics and their Implications on Search Engines, Almaden Research Center, Research Report, RJ 10199 (95073).

Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)
0-7695-2497-4/06 $20.00 © 2006 IEEE