

Extracting Semantically Related Queries By Exploiting User Session Information

Silviu Cucerzan and Eric Brill
Microsoft Research
Redmond, WA 98052, USA
[silviu;brill}@microsoft.com](mailto:{silviu;brill}@microsoft.com)

ABSTRACT

This paper presents a simple and very effective collaborative approach to generate semantically related queries to a user query by employing aggregated user session statistics, as captured by search engine query logs. We show empirical evidence that one of the main causes of the temporal correlation between semantically related queries, which was previously reported in the literature, is the fact that such queries are submitted by the same users in their search sessions. We also propose two evaluation methods that use real user queries from search engine query logs and WordNet data. To our knowledge, these represent the first automatic, non post-hoc, evaluation methods discussed in the literature that do not require the deployment of a system for generating semantically related queries with a commercial search engine. Finally, we discuss the computational performance of our approach and propose several directions for using semantically related queries.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Query formulation, search process, information filtering, clustering.*

General Terms

Algorithms, Experimentation, Languages.

Keywords

Search engine, query log analysis, user sessions.

1. INTRODUCTION

Web search engines have become a central entry point to the internet and a crucial factor in the experience of internet users, to the point that they are used for navigational and transactional needs even more than for informational needs, as shown by Broder [4] and Rose and Levinson [16]. As a consequence of their popularity, web search engines have added numerous new tools and features to help users with their internet experience, such as query spelling correction (e.g., Did You Mean – Google), web page translation (e.g., Babel Fish – AltaVista), query federation (e.g., Instant Answers – MSN Search), question answering (e.g., Encarta Answers – MSN Search), and direct navigation (e.g., I’m

Feeling Lucky – Google). Among such features, Yahoo (Also Try) and Lycos (Narrow Your Search) provide related queries to users. Yahoo provides query expansion, where the engine appears to provide the most popular queries that are superstrings of the target query, while Lycos provides either more specific or more general search choices, depending on the target query.

In this paper, we investigate an approach for generating semantically related queries to a target query and we focus our attention on the queries that are not superstrings of the target query. This type of related queries, with a lexical composition different than the composition of the target query, constitutes an interesting category for both research and applications, which has not been explored before. For an input query such as “*traffic*”, suggestions such as “*road conditions*” and “*steve winwood*”¹ may allow the users to derive searches with improved results or results closer to their search intent. Moreover, such suggestions can be used to cluster search results and even to derive search ontologies.

One way to generate semantically related queries, which we explore in this paper, is to use information extracted from previous *web search sessions* of users that submitted the same or similar queries. We define a session as a sequence of queries submitted to the search engine from the same browser window. While the information gathered from any single search session may not be reliable and contains lots of noise (because users may change their search focus in an unpredictable manner), we show that the aggregated information obtained from many user search sessions contains a tremendous amount of useful information. Instead of using temporal query patterns and expensive computational methods for retrieving the most similar temporal patterns (e.g., [5], [18]), we start from the premise that queries that follow each other in a substantial number of user search sessions are semantically related and we propose an *Occam’s razor’s* approach, of using directly the aggregated statistics over query sequence pairs to hypothesize semantically related queries. Although other definitions of query sequence pairs are possible, as discussed in Section 4.1, we focus mainly on queries that were sent in succession by a user to a search engine.²

In the absence of user studies to determine the usefulness of semantically related queries for web search, the previous work in this area has only presented evaluations on tens of hand picked examples. In this paper, we investigate two automatic ways to evaluate the performance of systems for generating semantically

¹ Steve Winwood has been the leader of the rock band Traffic.

² By storing only aggregated counts for pairs of queries that came in succession in search sessions, we also preserve better the anonymity of the users.

related queries, by using query log information and existing ontologies such as WordNet. We also discuss some possible uses of semantically related queries and explore the idea of building soft ontologies (i.e. concepts can be present in multiple positions with various probabilities) based on semantically related queries.

The remainder of the paper is organized as follows. In Section 2, we briefly review a number of relevant studies on query refinement and discovery of semantically similar queries. Section 3 presents the motivation of our approach in contrast with previous work. Section 4 describes the model based on user session statistics. In Section 5, we evaluate this model and compare our results with previously reported results. We propose methods of evaluating such systems by using query logs and information stored in ontologies such as WordNet. Details of our implementation of the system and computation performance are provided in Section 6. Section 7 discusses several directions of using the hypothesized semantically similar queries. Finally, Section 8 draws several conclusions and points to some further research directions.

2. RELATED WORK

There has been a considerable amount of work related to the discovery of semantically similar queries, both as a self-contained task and as a subtask for other tasks such as query expansion and clustering of search results. In one of the earliest studies in this area, Beeferman and Berger [1] propose an approach that exploits search click-through data, in which they represent the user queries and urls clicked by users as a bipartite graph and they apply an agglomerative clustering technique to identify related queries and URLs. Further, they suggest how to use the identified clusters to assist users in web search. In related work, Wen et al. [19] use lexical composition of the queries in conjunction with click-through data to determine alternate query forms to frequently asked questions for which sets of editorially checked relevant documents exist. In another study on using related queries, Daumé and Brill [8] cluster the documents retrieved by a search engine for a query based on alternative queries that returned common documents. Their working hypothesis is that queries that share a large number of common search results are semantically related and that the various semantic interpretations of a query can be hypothesized by partitioning the search result space. In a different study on improving the search relevance, Cui et al. [7] investigate the use on click-through information for query expansion. After showing that the lexical characteristics of the query space and the web document space are different, they investigate the mapping between query words and the words in the user-clicked documents in order to perform query expansion of user queries by using words characteristic to related web documents. In [2], Billerbeck et al. propose an effective method of obtaining query expansion terms from the past queries that retrieved the documents in the collection associated with a target query, reporting 26%-29% relative improvements over unexpanded retrieval on the TREC-9 and TREC-10 collections. Many other studies in the area of query expansion, query refinement, and semantically-similar query detection have proposed techniques based on various other sources of information, from pseudo-relevance-feedback (e.g., [15],[13]) and thesauri-based query expansion (e.g., [14]) to the use of pre-computed document abstracts (e.g., [3]) and anchor text information (e.g., [10]). Most of this work focused on bridging the informational need conveyed in a query to its lexical composition relied upon by search engines, rather than exploring the space of search concepts that are semantically and/or

ontologically related. The two studies that are most related to our current work are those published by Vlachos et al. [18] and Chien and Immorlica [5] on using query log statistics to hypothesize semantically similar queries. Vlachos et al. [18] build a time series for each query in which the elements are the daily frequencies of the query over a given period of time (three years) and they hypothesize semantically related queries by using the best Fourier coefficients as a lower bound on the Euclidean distance between the series. They show through examples that this technique is efficient especially for queries with similar annual burst patterns, such as “*Christmas*” and “*Rudolph the red nosed reindeer*”. Chien and Immorlica [5] explore the use of various time units (three hours, six hours, and twenty-four hours) over which they collect frequency information about queries in order to build the time series and they use the correlation coefficient between the frequency-based time series to hypothesize semantically related queries to any given query. They report that “for the top 100 most popular queries, for a weighted 70%, at least three of their top ten correlations were judged to be in fact semantically related” and they also present the most similar 10 queries produced by their system for a set of 17 queries (this set, as well as the set of queries reported by Beeferman and Berger pointed [1], together with the queries produced by our system, can be found in the Appendix).

3. MOTIVATION

Temporal correlation can be exploited successfully for some of the high-frequency queries, especially high-frequency queries with seasonal spikes (e.g., “*christmas*” and “*santa claus*”), as previously shown by Vlachos et al. [18], and event-driven high-frequency queries (e.g., “*scott peterson*” and “*peterson trial*”), as shown by Chien and Immorlica [5]. Nonetheless, it typically cannot handle many low-frequency queries (such as “*steel garage doors*” and “*garage door replacement*”); also, it is less useful for query pairs such as “*fishing rods*” and “*fly tying*”, even though they do have seasonal patterns, because of their low volume and because of the very large number of other queries with low volume with which they may be temporally correlated by chance. While we have not performed extensive experiments on using temporal correlation, we observed that this computational-costly technique also does not perform very well on many high-frequency queries without seasonal spikes or that do not refer to one event, such as “*britney spears*” and “*jennifer lopez*”.

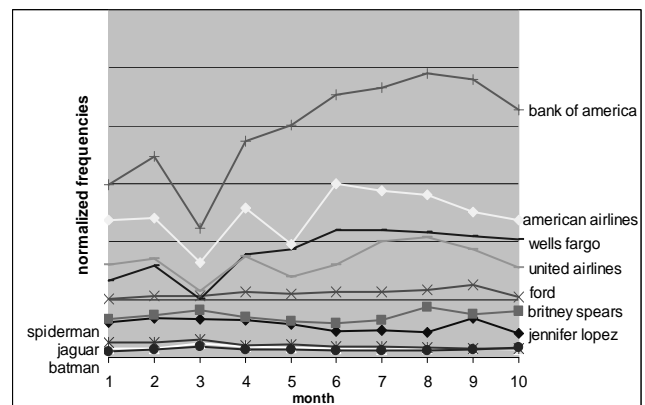


Figure 1. Relative monthly frequencies of ten queries (normalized by the total monthly volume of queries) over a 10-month period.

Table 1. Temporal correlation computed using normalized monthly query frequencies over 10 months.

Query	jennifer lopez	britney spears	jaguar	ford	spiderman	batman	bank of america	wells fargo	united airlines	american airlines
jennifer lopez	1.00	<i>-.01</i>	.48	<i>-.01</i>	.63	.16	<i>-.59</i>	<i>-.66</i>	<i>-.31</i>	<i>-.52</i>
britney spears	<i>-.01</i>	1.00	.24	.01	.00	.58	<i>-.08</i>	<i>-.18</i>	.05	<i>-.28</i>
jaguar	.48	.24	1.00	<i>-.19</i>	.75	.69	<i>-.71</i>	<i>-.71</i>	<i>-.69</i>	<i>-.72</i>
ford	<i>-.01</i>	.01	<i>-.19</i>	1.00	<i>-.53</i>	<i>-.23</i>	.71	.63	.59	.47
spiderman	.63	.00	.75	<i>-.53</i>	1.00	<i>0.35</i>	<i>-.92</i>	<i>-.92</i>	<i>-.64</i>	<i>-.65</i>
batman	.16	.58	.69	<i>-.23</i>	<i>0.35</i>	1.00	<i>-.46</i>	<i>-.45</i>	<i>-.58</i>	<i>-.64</i>
bank of america	<i>-.59</i>	<i>-.08</i>	<i>-.71</i>	.71	<i>-.92</i>	<i>-.46</i>	1.00	.98	.75	.75
wells fargo	<i>-.66</i>	<i>-.18</i>	<i>-.71</i>	.63	<i>-.92</i>	<i>-.45</i>	.98	1.00	.69	.77
united airlines	<i>-.31</i>	.05	<i>-.69</i>	.59	<i>-.64</i>	<i>-.58</i>	.75	.69	1.00	.82
american airlines	<i>-.52</i>	<i>-.28</i>	<i>-.72</i>	.47	<i>-.65</i>	<i>-.64</i>	.75	.77	.82	1.00

For example, Figure 1 shows the normalized frequencies for ten high-volume queries over a period of 10 months. We divided the monthly frequencies of each query type by the total volume of queries in each month to normalize out the effects of query volume differences over different months.

Table 1 shows the numerical values for temporal correlation between those queries, computed by using the monthly frequency time series. On each line, the values that are higher than the correlation coefficient for the semantically related query are in bold. While, as one would hope, “wells fargo” has the highest temporal correlation coefficient with “bank of america” in the set of investigated queries, “britney spears” is more strongly correlated with queries such as “jaguar”, “batman”, and “united airlines” than with “jennifer lopez”.

Therefore, it is natural to ask ourselves whether or not there is some hidden cause that makes temporal correlation work very well for some high-frequency queries that are not seasonal or event-driven, but performs rather poorly on others. As we show further, a very important quantitative component of the temporal correlation between queries such as “united airlines” and “american airlines” or “walmart” and “target” (one of the positive examples for temporal correlation used in previous work [5]) comes as a consequence of the fact that many users are querying both queries in the same search session.³ For example, Table 2 shows the most frequent queries that precede and follow the queries “walmart” and “target” in user sessions, as observed over a period of several weeks in the MSN Search query logs.⁴

Our experiments show that many of the semantically related queries hypothesized by using temporal correlation as semantically related to a target query are in fact queries that are among the top 10 most frequent queries that follow immediately the target query in user sessions (as shown in the Appendix).

³ Preliminary experiments also show the same users tend to re-query them over time, especially in the case of navigational queries.

⁴ We do not report the exact number in order to protect the MSN Search traffic volume information.

Table 2. Most frequent queries that precede and follow the queries “walmart” and “target” in user search sessions, and the corresponding precede and follow frequencies.

Target query: <i>walmart</i>			
Follows		Precedes	
Freq	Query	Freq	Query
3932	target	3599	target
2254	kmart	1639	kmart
1074	sears	1059	sears
1046	best buy	1024	best buy
868	toys r us	899	toys r us
523	circuit city	566	circuit city
371	toysrus	443	toysrus
357	lowes	384	home depot
331	home depot	310	sams club
326	sams club	306	lowes
Target query: <i>target</i>			
Follows		Follows	
Freq	Query	Freq	Query
3599	walmart	3932	walmart
1537	kmart	1314	kmart
878	sears	766	toys r us
790	best buy	741	best buy
751	toys r us	701	sears
433	kohls	456	wal mart
358	wal mart	381	circuit city
352	babies r us	379	wal-mart
350	circuit city	376	kohls
344	wal-mart	365	babies r us

Based on the observation that the temporal correlation of query frequencies over time is due in part to users that submit the related queries in the same search session, we now focus our attention to using a different type of query log evidence for generating semantically related queries: user search session statistics. In particular, we investigate how many times two queries appear adjacently in a user search session (i.e. they are queried in succession from the same browser window). Is it the case that, despite all the switches in search focus during user sessions, some queries appear very frequently in succession?

Figure 2 shows the distributions of the 100 queries that follow most frequently five of the ten target queries analyzed above over a period of several weeks. As it can be observed, these distributions of queries that follow are Zipfian [20], with several queries that follow extremely frequently the target queries and a large number of other queries that follow them a small number of times. Note that for queries such as “britney spears” and “united airlines”, even when using query log statistics collected only over several weeks, there are more than 50 queries that follow them at least 25 times in user sessions.

Moreover, an investigation of most frequent follow queries reveals that they can be judged, with very few exceptions, as being semantically related to the target queries that they follow.

These observations constitute the premises to hypothesize that queries that co-occur frequently in the same session are semantically related and thus, to attempt to use a collaborative technique to determine semantically related queries.

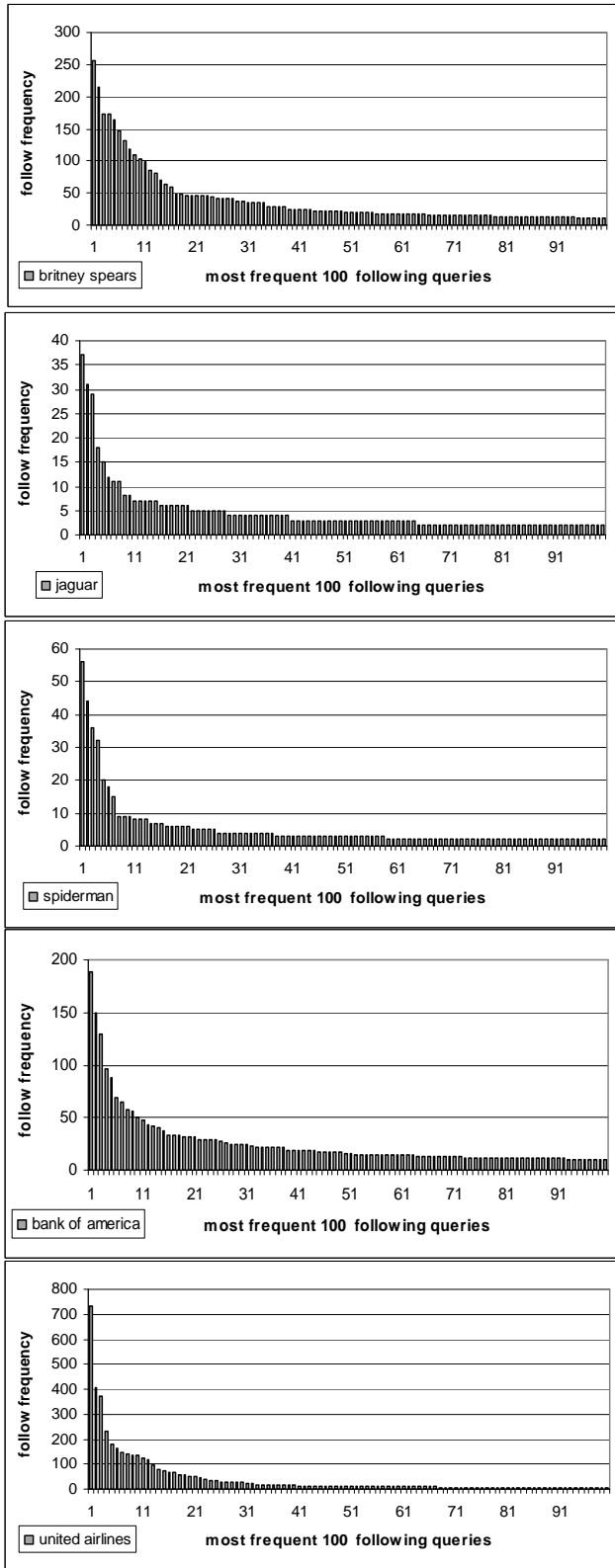


Figure 2. The frequencies of the top 100 follow queries over a period of several weeks for the queries “britney spears”, “jaguar”, “spiderman”, “bank of america”, and “united airlines”.

4. USER SESSION MODEL

Starting from the empirical evidence presented in the previous section, we propose an approach to generate semantically related queries based on query co-occurrence in user search sessions, which can be summarized as follows:

Collect co-occurrence statistics for all queries submitted by users over a long period of time. Use maximum likelihood estimation (MLE) to approximate the probability that a query q follows immediately another query p in a user search session and the probability of a query p to be sent by a user to a search engine:

$$P(q | p) = \frac{P(p, q)}{P(p)} \cong \frac{Freq(p, q)}{Freq(p)}$$

$$P(p) \cong \frac{Freq(p)}{\sum_{p' \in QLog} Freq(p')}$$

To generate semantically similar queries to a target query p , we first determine the queries $Q_p = \{q_1, q_2, \dots, q_N\}$ with the highest probability of following the query p . From this list, we eliminate the *stop queries*, which are queries that co-occur frequently with a large number of other queries. These queries can be obtained in a global way, by counting for each query the number of different *query types* (i.e. ignoring occurrence frequency) that it follows in user sessions and then hypothesizing the queries with very high counts as stop queries. A second approach is to discard from the list of queries Q_p as stop queries those queries q_i for which $P(q_i | p) / P(q_i)$ is low, which is equivalent to saying that they have low pointwise mutual information (PMI) with the target query. Note that, in this case, we cannot store a global list of stop queries and we have to compute them on the fly based on the target query.

While we expect that such an approach may work well on high-frequency queries, one important question is whether it can also achieve reasonable performance on the less frequent queries. In practice, a preliminary condition is to have stored *follow statistics* from previous query logs for a large subset of the queries received by a search engine. To determine whether or not this preliminary condition can be met, we computed the mean and standard deviation for the frequencies of the top follow queries for four different random sets, which were sampled from an aggregated query log with a cut-off frequency of 10 over a period of ten months. The follow statistics were collected from a non-overlapping query log over several weeks, which will be referred to as the *reference query log* henceforth. The two query logs were separated in time by more than six months, so that most of the event-driven queries in the two logs would be different.

First, we analyze 1000 queries sampled at random *by type*, which means that we disregard the number of instances of each query and give any two queries the same chance of being sampled, regardless of the number of instances they were sent to the search engine. From these 1000 queries, 723 appeared in the reference query log and 637 appeared in at least one session in which they were followed by another query. Figure 3 shows that on average, the most frequent follow query has an average frequency of 1.4, with a standard deviation of 2, the second most frequent follow query has a frequency close to 0.9, with a standard deviation of 1, and so on. While these first results may not show promise for reliably distinguishing information from noise in the sets of follow queries for most target queries, an empirical investigation of the follow queries shows that they could provide substantial information even for such small numbers.

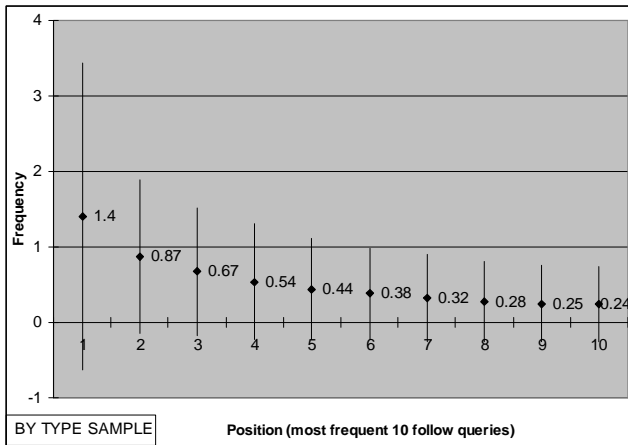


Figure 3. Mean and standard deviation of the frequencies of the top 10 follow queries over a period of several weeks for a random set of 1000 queries sampled by type from a previous 10-month query log

To exemplify, we present the data for three random queries from this first set: the query “ecosystems” is followed by “ecosystem”, “tropical rainforest”, “savannah desert”, “animals of the deciduous forest”, and “ask jeeves”; the query “wooden high chairs” is followed by “baby nursery ideas”, “community playthings”, “darling woodcraft”, “one step ahead”, “wal mart”, and “wooden baby cribs”; the query “vietnamization” is followed by “nixon doctrine”, “communism”, “ending of vietnam war”, “khmer rouge”, and “munich accords”. Most of these follow queries are semantically related with the target query, despite the fact that the follow counts are extremely low.

The queries in this first sample have an average length of 2.75 words. As expected, we observed that the shorter the query the more follow queries with higher counts exist. To validate this observation, we analyzed another random sample of 1000 one-word queries, sampled again by type. Of these, 889 queries appear in the reference query log and 801 appear in at least one session in which they are followed by another query. The statistics for this sample confirmed our hypothesis: the average frequency of the most frequent follow query increases to 2.75, the second most frequent and third most frequent being 1.52 and 1.18 respectively.

While these statistics show that there might not be enough statistical evidence in a query log collected over only a few weeks to reliably hypothesize semantically related queries based on user sessions for the average query type, we show further that such statistics are appropriate when we take into consideration the query traffic and perform the sampling *by token*. This means that a search engine could suggest semantically related queries most of the time based on the proposed approach.

Figure 4 shows that, when we use a 1000-count query set randomly sampled by token (i.e. each query has a chance of being selected in the sample proportional to its frequency), the average frequency of the queries most frequently following the target queries is no less than 442 when no word-length restriction is imposed, and 985 of the sample queries appeared in the reference log. In a similar experiment, by sampling one-word queries by token, the average frequency of the most frequently follow query is 1126, and 999 of the 1000 sample queries appeared in the reference log. These statistics show that the proposed approach of using the follow queries has a great potential for being used in practice by a search engine.

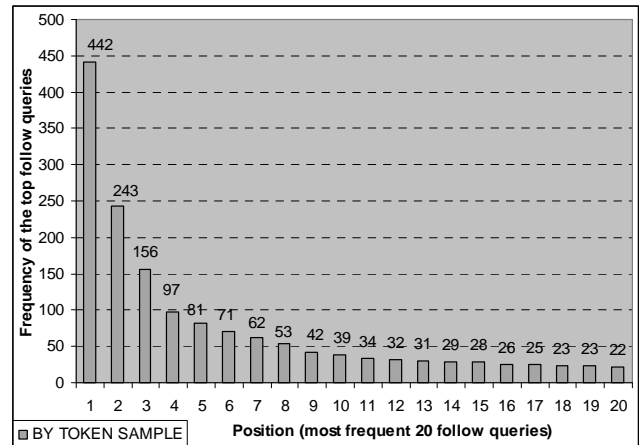


Figure 4. Average frequencies of the top 20 follow queries over a period of several weeks for a random set of 1000 queries sampled by token from a previous 10-month query log

4.1 Further Analysis

4.1.1 Aggregated follow and precede statistics

One variation of the proposed approach is to multiply the follow and precede counts and rank the semantically similar queries by this score. For example, the most frequent follow queries for “space shuttle” are “space shuttle pictures”, “space shuttle columbia”, “hummer”, “humvee”, “nasa”, “space shuttle missions”, “enterprise”, and “kennedy space center”, while the most frequent queries that precede it are “nasa”, “space shuttle columbia”, “space”, “shakespeare”, “space ship”, “apollo 13”, “columbia”, and “mars rover”. When we multiply the follow and precede counts, we hypothesize the most similar five queries to be “nasa”, “space shuttle columbia”, “columbia”, “moon landing”, and “space exploration”. The queries for which the multiplication result is zero (such as “hummer”) can be filtered out. This approach can produce higher quality suggestions and is also more robust to misspellings of a target query (which tend to precede it but not follow it) and stop queries (which tend to follow more often than precede other queries, based on how they were defined).

4.1.2 Types of semantically similar queries

As it can be observed in the example in the previous section, several of the most frequent follow queries are in fact superstrings of the target query. In general, we distinguish three main types of suggested queries:

- queries that are a substring of the target query,
- queries that are a superstring of the target query, and
- queries that are neither of these.

The queries of type (a) and (b) can be exploited for various purposes, such as to determine collocations and noun phrase boundaries or to perform word sense disambiguation. The superstring queries may be extremely valuable for users of a search engine, as they represent refinements that may disambiguate the target query, typically formulated by users that were not satisfying with the results returned by the search engine for the target query. The statistics on the substring queries can be used to split the target query into terms of interests or typical collocations.

However, we consider the third category of queries (c) to be the most interesting, as it allows us to discover distinct concepts that

are related to each other rather than disambiguating attributes for one lexical item. In the remainder of this paper, we will only discuss the production of queries that have a different lexical composition than the target query’s lexical composition. To restrict the space of suggestions to only such queries, we devise the system to eliminate from the list of follow queries not only substrings and superstring, but also *approximate duplicates*, as discussed further.

4.1.3 Approximate duplicate removal

In order to remove approximate duplicates such as “*bank of america*” and “*bankofamerica*”, or “*dog photos*”, “*dog photo*” and “*photos of dogs*” from the list of semantically similar queries, the system we implemented compares each hypothesized query with the target query and all of the already hypothesized queries by checking whether:

- a) the queries contain the same non-stop words (this can be easily done by sorting alphabetically the non-stop words of the queries),
- b) the queries contain the same non-stop word stems;
- c) the concatenated forms of the queries (i.e. obtained by removing the word delimiters such as spaces and hyphens) are the same.

When eliminating approximate duplicates, we keep as a candidate the query form that follows most frequently the target query.

4.1.4 Back-off for long queries

For long queries, for which there is a lack of session information (e.g. “*hard disk case*”), we employ a back-off procedure, in which the leftmost or rightmost words are removed iteratively until we obtain the longest possible query with a log frequency above a certain minimum threshold and for which the number of query extensions is below a certain maximum threshold (e.g. “*hard disk*”). We employ the former restriction to avoid backing off to rare queries, for which reliable follow statistics may not exist, and the latter to avoid as much as possible backing off to underspecified queries, e.g., from a query such as “*john sundermeyer*” to the unrelated query “*john*”.

5. EVALUATION

5.1 Post-hoc Evaluation

As Beeferman and Berger pointed out in [1], it may be difficult to judge the hypothesized semantically related queries on the basis whether they “make sense together” with the target query without having a specific application in mind. If the suggested queries are intended to be presented directly to the users of a search engine then the best possible measurement is the click-through rate (which by itself, is a very rough measure of usefulness if not observed over a long period of time). Nonetheless, in the absence of click-through information, and because our purpose is to evaluate one particular class of related queries, namely those that have a lexicon composition different than the target query, we start this section with a comparison of our approach with the only other post-hoc evaluation of semantically related queries mentioned in literature. Chien and Immorlica [5] performed a post-hoc evaluation of their system on the 100 most popular queries, reporting that for a weighted 70%, at least three of their top ten correlations were judged to be in fact semantically related. We performed a similar experiment, generating and judging the top three hypothesized queries for the 100 most popular queries. Two annotators analyzed independently the query set and judged that 94%, respectively 98% of the time, at least one of the top three hypothesized queries is semantically similar to the target

query and that 60%, respectively 72% of the time, all three suggested queries are semantically similar to the target query.

5.2 Query-log-based Evaluation

Because even a post-hoc direct evaluation of the quality of semantically related queries hypothesized is very costly, we investigated an indirect evaluation methodology which uses the most frequent *query extensions*, as observed in the search query log. For example, the query “*ravens*” has the most common query extensions, as observed in our reference query log, “*football*” with a frequency of 31 (i.e. the query “*ravens football*” was sent by users 31 times to the search engine), “*tickets*” (27), “*stadium*” (19), “*psl*” (15), “*cheerleaders*” (14), etc., while the query “*bears*” has the most common extensions “*furniture*” (32), “*pictures*” (20), “*tickets*” (19), “*den*” (14), “*football*” (12), etc.

Ideally, one would generate semantically similar queries to a target query by computing the queries in the query log space with the most similar extensions. This is not feasible because the candidate space and the attribute space are huge. Nonetheless, we can use the extensions to evaluate the performance of methods for generating semantically similar queries as we propose further.

Formally, for a target query p for which a set of queries Q are hypothesized as semantically related, we measure the distributional similarity between the query extensions for p (denoted by E_p) and the aggregated query extensions of the queries in Q (denoted by E_Q).

A common way to measure the difference between two distributions is relative entropy, also known as the *Kullback-Leibler divergence*,

$$\text{KL}(E_p \parallel E_Q) = \sum_w E_p(w) \cdot \log \frac{E_p(w)}{E_Q(w)},$$

which represents the average number of bits wasted by encoding events drawn from a distribution E_p using as model the distribution E_Q . In order to avoid dealing with zero-values, for which the logarithm is undefined, we use the symmetric relative entropy, also known as *Jensen-Shannon divergence*: $\text{JS}(E_p, E_Q)$

$$= \frac{1}{2} (\text{KL}(E_p \parallel \text{avg}(E_p, E_Q)) + \text{KL}(E_Q \parallel \text{avg}(E_p, E_Q)))$$

Because the value of the JS-divergence would be difficult to interpret in the absence of a baseline, we compute a reference value as follows: for each target query p for which we are able to hypothesize a set Q of semantically similar queries, we select at random a set of queries Q_{RAND} such that $|Q_{\text{RAND}}| = |Q|$. We then compare $\text{JS}(E_p, E_Q)$ with the baseline $\text{JS}(E_p, E_{Q_{\text{RAND}}})$.

Table 3. The mean and standard deviation of the JS-divergence between query extension distributions (the lower the value the more similar two distributions are)

Random Query Set	Baseline $\text{JS}(E_p, E_{Q_{\text{RAND}}})$		$\text{JS}(E_p, E_Q)$	
	Avg	Std dev	Avg	Std dev
Distributions over query extension				
Any length queries	0.97	0.04	0.72	0.18
One-word queries	0.97	0.04	0.77	0.16
Distributions over query extensions’ component words				
Any length queries	0.94	0.05	0.67	0.18
One-word queries	0.95	0.05	0.72	0.16

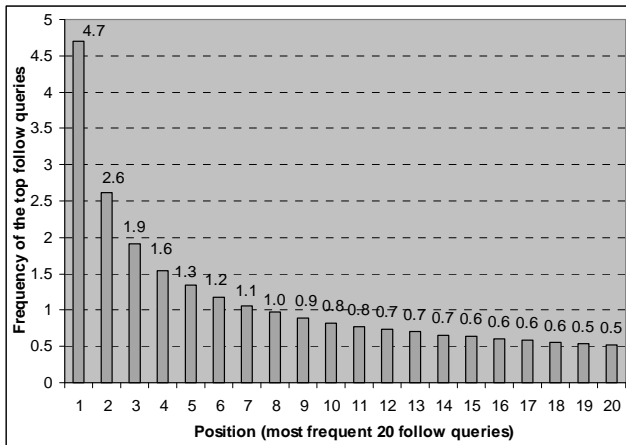


Figure 5. Frequencies of the top 20 follow queries over a period of several weeks averaged over all noun phrases in WordNet 2.1

Table 3 shows the mean and standard deviation for the query-extension-based JS-divergence values for the two sets of random queries selected by token that were analyzed in Section 4. To avoid counting zero values, we used only target queries for which semantically similar queries can be hypothesized. The values obtained are extremely promising, especially considering the fact that pairs of queries with very similar extension distributions and thus, very low JS-divergence values, are extremely rare (e.g., the JS-divergence for “ravens” and “bears” when considering only the distributions over the five most popular extension is 0.59).

5.3 WordNet-based Evaluation

We now propose another evaluation of systems that hypothesize semantically related queries by using WordNet 2.1, the online lexical reference system for English developed at Princeton University [12]. This collection contains nouns, verbs, adjectives, and adverbs organized into synsets, each representing one underlying lexical concept. In our evaluation, we used all 67,504 unique terms in the WordNet 2.1 noun file as search queries. Of these, 48,674 appeared as logged user queries and 45,259 appeared in at least one session with a follow query. On average, the most frequent follow query has a count of 4.7 (i.e. it was submitted in 4.7 user sessions immediately after the target query) and the second most frequent has a count of 2.6. Statistics up to the 20th most frequent follow query are shown in Figure 5. Note that on average, there are 3 queries which follow twice or more the WordNet terms. For example, “face lift” is followed by “liposuction” three times, “cosmetic surgery”, “scar removal”, “face lifts”, “facelift”, and “non surgical face lift” twice, etc.

While these results show there is promise in using the follow queries for the information contained in WordNet, they do not indicate the quality of these follow queries. To quantify their quality, we tested how many times the follow queries were directly related to the input terms according to the WordNet hierarchy, as hyponyms, hypernyms, or synonyms.

We obtained that 1,746 of the follow queries were hyponyms of the target query they followed, 1,659 were hypernyms of the target query, and 1,186 were synonyms of the target query.

Also, there are an additional 7,420 instances of hyponyms of the target query contained, 63,748 instances of hypernyms, and 38,429 instances of synonyms as parts the follow queries.

Table 4. Number of follow queries that are or contain a hyponym, hypernym, or synonym of the target query by frequency

Type	Freq	≥ 50	≥ 25	≥ 10	≥ 5	≥ 2	= 1
Exact hyponym		7	15	91	206	725	702
Contains hyponym		11	43	166	426	1977	4797
Exact hypernym		6	8	49	110	533	953
Contains hypernym		28	93	579	1693	9983	51372
Exact synonym		4	7	32	77	379	685
Contains synonym		15	58	292	800	5380	31882

6. IMPLEMENTATION DETAILS

To implement a basic system based on the algorithm described in Section 4, one has to store a histogram of pairs of queries that were sent in succession to a search engine by users over a certain period of time. Based on space limitations and the reliability threshold, a certain frequency cut-off value can be assumed.

In our particular implementation, we stored the query log statistics over several weeks into a SQL database with multiple indexes that allow fast retrieving of any number of the most frequent precede and follow queries, on a dual 2.4 GHz Intel Xeon server with 3 GB of RAM. On a separate identical server, we created a web service that accesses the database and computes the semantically similar queries to input queries. This system is able to handle over 25 input queries per second. In our experiments, the system accomplished the task of computing exhaustive lists of similar queries for 4 random sets of 1000 queries discussed in Section 4 (which will be made available to the academic community), including the I/O operations, in 14, 23, 28, and 29 seconds (on average, over 47 input queries per second). On the WordNet terms, the system averaged 26 queries per second.

In order to eliminate misspelled queries from the list of semantically similar queries suggested by our system, we employed a spell checker trained on search query logs, similar to that proposed by Cucerzan and Brill [5], which also blocks queries containing offensive terms. All examples used in this paper were validated by this spelling correction system (thus, in some cases, more popular misspellings or offensive queries than the queries shown may exist).

7. USING THE RELATED QUERIES

For a query such as “britney spears”, our system retrieves, based on user session statistics, the related queries “christina aguilera”, “jennifer lopez”, “barbie”, “justin timberlake”, “jessica simpson”, “madonna”, “jason allen alexander”, “paris hilton”, “avril lavigne”, and “eminem”. Because users who queried for Britney Spears in the past were also interested in these other celebrities, as the follow statistics indicate, providing links to the search results for these celebrities may be valuable to the search engine user.

One natural question is whether or not the hypothesized semantically related queries can be used for building web search ontologies and characterizing similar search terms/concepts by similar attributes. In an attempt to answer this question, we performed a preliminary investigation of obtaining such attributes automatically, in a manner inspired by the evaluation experiment that uses query extensions presented in Section 5.2, in which we built a system that hypothesizes attributes for a search concept (target query) based on the query extensions that are common to at least a certain fraction of the semantically similar queries. Although we have not performed a thorough evaluation of this system yet, we observed that such a system was able to put

Table 5. Popular attributes for the conceptual class of a target query, as extracted by using the query extensions common to the majority of the hypothesized semantically similar queries.

Query	Top 10 semantically-related queries	Popular concepts
britney spears	christina aguilera, jennifer lopez, barbie, justin timberlake, jessica simpson, madonna, jason allen alexander, paris hilton, avril lavigne, eminem	links, galleries, fan club, pictures, facts, music, backgrounds, wallpaper, gossip, song lyrics, fashion, music video, clothes, posters, concert, official site, screensavers, news, images, quotes, clothing, mp3, picture gallery, buddy icons, downloads, birthday, official website, height, profile, discography, pic, movie, bio, photographs, fan site, album
bank of america	wells fargo, american express, capital one, washington mutual, bank one, wachovia, sprint pcs, citibank, providian, mbna	financial services, mastercard, bill pay, card services, home, credit, credit card, rewards, mortgage, auto loans, visa card, on line banking, human resources, homepage, employment, corporation, locations, jobs, careers, home page, customer service, bank, accounts, insurance, gift cards
united airlines	american airlines, delta airlines, southwest airlines, northwest airlines, continental airlines, expedia, travelocity, us airways, orbitz, alaska airlines	coupons, airfares, flight, vacations, cargo, schedules, home, mexico, telephone number, cargo tracking, information, confirmation, check in, flight info, credit cards, flight tracker, flight confirmation, arrivals, visa, flight schedule, destinations, flight information, reservations, homepage, employment, vacation packages, phone number, frequent flyer, credit union, ceo, jobs, careers, mastercard, customer service, employees, flight status, arrival times, history, flight arrivals, magazine
jaguar	ford, bmw, mercedes, audi, aston martin, land rover, lexus, volvo, cadillac, mercedes benz	performance, club, forums, auto parts, performance parts, body kits, models, wallpaper, mexico, tires, body parts, recalls, suv, italia, cars, dealer, accessories, engines, floor mats, used, for sale, car dealers, used cars, convertible, leasing, car parts, merchandise, history, lease, wheels, automobiles, tuning, dealership, repair, occasion, service, used parts, atlanta, rims, spares, racing, insurance
spiderman	batman, hulk, superman, kirsten dunst, disney, harry potter, x-men, marvel comics, power rangers, barbie	action figures, movies, store, cartoons, buddy icons, picture, clip art, backgrounds, clipart, wallpaper, soundtrack, merchandise, clothes, posters, desktop wallpaper, cast, books, gallery, coloring pages, images, clothing, drawings, toys, bedding, downloads, t-shirts, screensavers, costume, coloring book, games online, cards, watches, characters, history, screen savers, dvd

forward extremely high quality attributes for the overwhelming majority of concepts investigated. For example, among the query extensions that are common to at least 70% of the semantically related queries hypothesized for “*britney spears*”, we find “*fan club*”, “*pictures*”, “*music*”, “*gossip*”, “*song lyrics*”, “*fashion*”, “*music video*”, “*posters*”, “*concert*”, “*news*”, “*mp3*”, “*downloads*”, “*birthday*”, “*discography*”, “*bio*”, “*fan site*”, and “*album*” (a complete list ordered by frequency is shown in Table 5). All these represent good indicators that the queried term is a music artist. Similarly, for a query such as “*bank of america*”, we identify as query extensions common to at least 70% of the semantically related queries concepts such as “*financial services*”, “*bill pay*”, “*card services*”, “*credit card*”, “*rewards*”, “*mortgage*”, “*auto loans*”, and “*on line banking*”, all being very good quality attributes for financial institutions.

We also observed that, while the query extensions common to most of the semantically related queries can be used to find the place of a target query in a hierarchy of search concepts, high frequency extensions of the target query that do not appear as extensions for the related queries (e.g. for “*britney spears*”: “*marriage*”, “*oops*”, “*wedding*”) can be used as distinguishing attributes that particularize the concept in the hierarchy.

8. CONCLUSION AND FUTURE WORK

We presented an efficient and effective approach to generate semantically related user queries, based on aggregated user session information statistics. Such an approach can be employed to address two of the major problems in web search: mapping a

search intent to an appropriate query form and empowering the users to discover and explore topics related to those they are querying about.

We showed that this collaborative approach for generating semantically related queries competes successfully with temporal correlation approaches that were previously presented in the literature. We also proposed two new methods for evaluating the quality of semantically related queries produced by a system and we evaluated the proposed approach on real user queries from the MSN Search logs and in conjunction with the WordNet ontology.

There are several research directions for both improving the quality of semantically related queries and using them to address other search problems. One direction for which we plan to do an in-depth investigation is the personalization of semantically related query suggestions, by hypothesizing topic labels for previous user sessions offline and using a collaborative filtering technique to label user search sessions in progress. As discussed in Section 7, another important direction of future research encompasses the building of soft ontologies based on query attributes determined automatically by using the semantically related queries, for which we plan to do a full evaluation and investigate practical scenarios for using it to improve the user’s experience in web search.

9. ACKNOWLEDGMENTS

We thank MSN Search for providing access to their query logs and allowing us to publish the results of our evaluations.

10. REFERENCES

- [1] Beeferman, D. and A. L. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of International KDD Conference 2000*, pp. 407-416.
- [2] Billerbeck, B., Scholer, F., Williams, H. E., Zobel, J. Query Expansion using Associated Queries. In *Proceedings of CIKM 2003*, pp. 2-9.
- [3] Billerbeck, B. and J. Zobel. Techniques for efficient query expansion. In *Proceedings of the String Processing and Information Retrieval Symposium 2004*, pp. 30-42.
- [4] Broder, A. Taxonomy of Web Search, *SIGIR Forum 2002*, 36(2).
- [5] Chien, S. and N. Immerlica. Semantic Similarity Between Search Engine Queries Using Temporal Correlation. In *Proceedings of the 14th International Conference on WWW 2005*, pp. 2-11.
- [6] Cucerzan, S. and E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2004*, pp. 293-300.
- [7] Cui, H., J. R. Wen, J. Y. Nie, and W. Y. Ma. Probabilistic Query Expansion using Query Logs. In *Proceedings of 11th International Conference on WWW 2002*, pp. 325-332.
- [8] Daumé, H. and E. Brill. Web search intent induction via automatic query reformulation. In *Proceedings of the HLT/NAACL 2004 Conference*, pp. 49-52.
- [9] Fonseca, B. M., P. Golgher, B. Póssas, B. Ribeiro-Neto, and N. Ziviani. Concept-Based Interactive Query Expansion, In *Proceedings of CIKM 2005*, pp. 696-703.
- [10] Kraft, R. and J. Y. Zien. Mining anchor text for query refinement. In *Proceedings of 13th International Conference on WWW 2004*, pp. 666-674.
- [11] Lee, L. Measures of distributional similarity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics 1999*, pp. 25-32.
- [12] Miller, G. A. WordNet: an online lexical database. In the *International Journal of Lexicography*, 3(4), 1990 (revised in 1993), pp. 245-264.
- [13] Mitra, M., A. Singhal., and C. Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st International ACM SIGIR Conference 1998*, pp. 206-214.
- [14] Qiu, Y. and H. Frei. Concept-based Query Expansion. In *Proceedings of the 16th International ACM SIGIR Conference 1993*, pp. 160-169.
- [15] Robertson, S. E. and K. Sparck Jones. Relevance weighting of search terms, In *Journal of the American Society of Information Science*, Vol. 27, May-June 1976, pp. 129-146.
- [16] Rose, D. E. and D. Levinson. Understanding User Goals in Web Search. In *Proceedings of 13th International Conference on WWW 2004*, pp. 13-19.
- [17] Spink, A. and H. C. Ozmutlu. What do people ask for on the web and how do they ask it: Ask Jeeves query analysis. In *Proceedings of ASIST 2001*, pp. 545-554.
- [18] Vlachos, M., C. Meek, Z. Vagena, and D. Gunopulos. Identification of Similarities, Periodicities & Bursts for Online Search Queries. In *Proceedings of SIGMOD 2004*, pp. 131-142.
- [19] Wen, J. R., J. Y. Nie, and H. J. Zhang. Clustering user queries of a search engine. In *Proceedings of 10th International Conference on WWW 2001*, pp. 162-168.
- [20] Zipf, G. K. *Selective Studies and the Principle of Relative frequency in Language*. Harvard University Press, 1932.

**Appendix. Comparison of the proposed system’s output and two previously reported systems on the previously published query sets.
In each case, the queries suggested by both the proposed system and a previous system are in bold.**

Target Query	Semantically similar queries that are not superstrings of the target query, as hypothesized based on user search sessions	Semantically similar queries reported by Chien and Immerlica [5] for a system using temporal correlation of query occurrences
income tax	revenue canada, turbo tax, irs, h&r block , e file, government of canada, national insurance, quick tax	tax returns, h&r block , federal income tax, taxact, efile
walmart	target , kmart, sears , best buy , toys r us, circuit city , lowes, home depot, sams club , costco	target , petsmart, circuit city , bed bath and beyond, sam’s club , best buy , sears , realtor.com
alice walker	maya angelou , gwendolyn brooks , toni cade bambara, langston hughes , nikki giovanni, toni morrison	michael jordan, martin luther king jr, jackie robinson, ella fitzgerald, malcom x, frederick douglas, maya angelou , langston hughes , gwendolyn brooks
greeting cards	birthday cards, e-cards, christmas cards, free e-cards , american greetings , hallmark, blue mountain cards, 123 greetings, email cards , invitations	free e-cards , egreetings.com, free cards, american greetings , bluemountain.com, yahoo cards, email cards , msn greetings
scott peterson	laci peterson , michael jackson, kobe bryant, dru sjodin, amber frey, court tv , elizabeth smart, sharon rocha, cnn, national enquirer	court tv , laci peterson , dictionary, modesto bee, scott peterson trial, peterson trial, washington mutual bank, free clip art, thesaurus
superbowl	nfl, southwest airlines, ticketmaster, cbs, expedia, reliant stadium, cnn, espn, football, mapquest	superbowl commercials, superbowl halftime show, superbowl xxxvii, superbowl 2004, nfl halftime show, superbowl-ads.com, janet jackson, pepsi commercial, tom brady
weather	news, mapquest, cnn, espn, road conditions, expedia, horoscope, nfl, noaa, yellow pages	weather channel, national weather service, weather underground, intellicast.com, accu weather, twc, nbc6.net, click10, msn weather
ajc	cnn, wsbtv, espn, usa today, monster, athens banner herald, gwinnett daily post, msnbc, nydailynews, atlanta falcons	cincinnati enquirer, times union, boston herald, baltimore sun, pittsburgh post gazette, bb&t, ny daily news, detroit free press, washingtonpost.com
bankone	bank of america , capital one, mbna, american express, discover, cardmemberservices, citibank, first usa, husker, providian	fleet, charter one bank, usajobs, wachovia bank, chevy chase bank, bankofamerica , usaa, air tran, amsouth
cnn	fox news , msnbc , bbc, espn, usa today, abc, cbs news , abc news , nasa, washington post	fox news , cbs news , drudge, abc news , msnbc , grainger, wpvi, aol, www.ups.com
dictionary	thesaurus , encyclopedia, ask jeeves, encarta, spelling, spell check, mapquest, translation, translator, bible	websters dictionary, thesaurus , free translation, dictionary.com, register to vote, spanish dictionary, free clip art
disney	barbie, nick, cartoon network , mickey mouse, winnie the pooh, nick jr, nemo, nickelodeon, expedia, universal studios	barbie.com, postopia.com, noggin.com, cartoon network , neopets.com, pbs kids, nickjr, bratz.com, yugioh.com
movies	blockbuster, music, fandango, lord of the rings, amc , cinema, games, entertainment, theatres, films	amc , ravemotionpictures.com, movie listings, movies.com, www.movifone.com, regalcinema.com, local movies, harkins.com, www.movietime.com
priceline	expedia, orbitz, hotwire , travelocity, cheap tickets , southwest, southwest airlines , airline tickets, northwest airlines, united airlines	cheap tickets , hotwire , www.orbitz.com, hotels.com, boat trader, southwest airlines , www.costco.com, niagara falls, ata airlines
sears	walmart , home depot, target, lowes, best buy, jc penney, circuit city , jc penny, kmart, kohls	barnes and noble, walmart , jcpenny, dell.com, comp usa, circuit city , u haul, office max, lowe’s
southwest airlines	american airlines , delta airlines, expedia , united airlines , orbitz , northwest airlines, travelocity , continental airlines, america west airlines, alaska airlines	orbitz , travelocity , united airlines , american airlines , frontier airlines, expedia , www.fafsa.ed.gov, cheap tickets, netflix
yellow pages	white pages, mapquest, maps, phone book, driving directions, phone numbers, zip codes , area codes, telephone directory, 411	zip codes , msn yellow pages, www.whitepages.com, yahoo yellow pages, us postal service, switchboard, social security, federal express, anywho.com
Target Query	Semantically similar queries that are not superstrings of the target query, as hypothesized based on user search sessions	Semantically similar queries reported by Beeferman and Berger [1] for a system using an agglomerative clustering technique
airline reservations	southwest airlines, delta airlines, expedia, airline tickets, american airlines, continental airlines, united airlines	travel agencies, plane tickets, travel agency, low airfare, hotel reservations, hotels carribean, travelocity, saber.com
bibliofind	amazon, book finder	rare books, used books, books used, out of print botany books, books antique, rare books’
e-mail cards	hallmark, blue mountain, blue mountain cards, email birthday cards, greeting cards	talk city, swan backgrounds, insurance seminar, creatacard, postnet.com, greetingcards, 4anything.com, lariam discussion groups
american airlines	united airlines, delta airlines, southwest airlines, continental airlines, expedia, northwest airlines, travelocity, orbitz, jet blue, cheap tickets	ww.aa.com, aa.com, air fairs, american airline, ‘american airlines’, american air, amr.com, american eagle airlines
amtrak	greyhound, mapquest, expedia, greyhound bus, southwest airlines, travelocity, orbitz, delta, cheap tickets, metrolink	amtrak schedules, amtrack, www.amtrack.com, amtrack reservations, amtrack.com, train amtrak, amtrack train schedule, train amtrack
gameboy emulator	gameboy roms, gameboy advance emulator, gameboy color emulator, gameboy download	emulator, nintendo roms, nes roms

