Keyword Generation for Search Engine Advertising

Amruta Joshi[1]
Department of Computer Science
Stanford University, USA
amrutaj@cs.stanford.edu

Rajeev Motwani[2]
Department of Computer Science
Stanford University, USA
rajeev@cs.stanford.edu

Abstract

Keyword[3] generation for search engine advertising is an important problem for sponsored search or paid-placement advertising. A recent strategy in this area is bidding on nonobvious yet relevant words, which are economically more viable. Targeting many such nonobvious words lowers the advertising cost, while delivering the same click volume as expensive words. Generating the right nonobvious yet relevant keywords is a challenging task. The challenge lies in not only finding relevant words, but also in finding many such words. In this paper, we present TermsNet, a novel approach to this problem. This approach leverages search engines to determine relevance between terms and captures their semantic relationships as a directed graph. By observing the neighbors of a term in such a graph, we generate the common as well as the nonobvious keywords related to a term.

1. Introduction

Search engine advertising has exploded in popularity over the past few years. With the benefits of the targeted audience, low cost per ad and reach of about 80% internet users, pay-per-click advertising is one of the most popular forms of online advertising today. In pay-per-click advertising, the ad is placed alongside search results with some fee for each click on the ad. Search engines set advertisers against each other in auction-style bidding for the highest ad placement positions on search result pages. The cost of the top position depends greatly upon the keyword you are bidding for. For example, the #1 ad ranking for the term 'hawaii vacation' will cost around $3 per click. Whereas, you could get the #1 position for the term 'kauai trip' for a mere 5 cents per click. The latter does not produce as much traffic as the former, but it is more economical. If you bid on a large number of these low-traffic keywords, the combined traffic from them adds up to the level produced by a popular keyword, at a fraction of the cost. Besides, the traffic received is targeted better and will typically result in a better clicks-per-sale rate. It is important to find out new alternative keywords, relevant to the base query, but nonobvious in nature, so that you face little competition from other advertisers. Thus, keywords generation for search engine advertising, also known as keyword research, is an important problem for sponsored searches. The objective is to generate, with good precision and recall, large number of terms that are highly relevant yet nonobvious to the given input keyword.

This paper proposes a novel graph-based technique called TermsNet to identify relevant yet nonobvious terms and their semantic associations. We present results of this technique applied to keyword research. However, we would like to point out that TermsNet is a general technique and can be extended to other applications such as documents matching, term clustering, study of word relationships.

We make the following salient contributions in this paper. We introduce the notion of directed relevance, i.e., relevance of keyword A to keyword B is independent of relevance of B to A. This is the key idea for exploring nonobvious relevance relationships. We propose TermsNet, a novel graph-based approach to keyword research. We define a new measure called nonobviousness, which is one of the unaddressed needs of keyword research. We provide experimental results and evaluation for TermsNet and other available tools for keyword research. To the best of our knowledge, no previous literature provides a study and/or comparative analysis of these.

The paper is organized as follows - Section 2 discusses other available techniques. Section 3 explains the proposed technique. Section 4 describes the experiments. Section 5 presents evaluation and results. Section 6 closes with conclusions & future work.

---

[1] Now working at Yahoo! Research Labs., CA, USA

[2] Supported in part by NSF Grant ITR-0331640, TRUST, and a grant from Media-X

[3] In this paper, the term 'keyword' to refers to words, phrases and query terms in general.

2. Related Techniques

In this section, we summarize the techniques used by other available tools for keyword research, specifically, meta-tag spiders, iterative query expansion, proximity-based searches, query-log and advertiser-log mining.

Many high ranked websites, using search engine optimization techniques, include relevant keywords in their meta-tags. A meta-tag spider queries search engine for seed keyword and extracts meta-tag words from these highly ranked webpages. Although there is no guarantee to find good keywords, these meta-tags open valuable directions for expansion. Popular online tools like Wordtracker [7] use meta-tag spidering for keyword suggestion. Another approach to extract related words is to use the Metacrawler Search Network's related keyword lists. Search engines maintain a list of few related keywords used for query expansion. To gather more words, current tools re-spider the first list of resulting keywords. This gives popular keywords closely related to the base keyword, but the number of relevant keywords generated is still low.

Proximity-based tools issue queries to a search engine to get highly ranked webpages for the seed keyword and expand the seed with words found in its proximity. For example for the seed keyword ‘hawaii vacations’, this tool will find keywords like: ‘hawaii family vacations’, ‘discount hawaii vacations’, etc. Though this tool finds a large number of keywords, it cannot find relevant keywords not containing the exact seed query words. The Google Adwords Tool [2] relies on query log mining for keyword generation. In Specific Matches, it presents frequent queries that contain the entire search term. Similarly, Overture’s Keyword Selection Tool lists frequent queries of recent past containing the seed terms. Both these techniques suffer from drawbacks like proximity-based searches, i.e., failure to generate relevant keywords not containing search terms. To generate Additional Keywords, Adwords mines advertiser logs. When searching for keyword ‘A’ to advertise, it presents other keywords which were searched for by other advertisers searching for ‘A’, i.e., it exploits co-occurrence relationships in advertiser query logs. Though this generates a large number of keywords, they are not always relevant. Also, keywords generated by this technique are limited to those words that occur frequently in advertiser search logs. Such frequent words have a good chance of being among expensive keywords, as they are already popular in the advertising community.

The existing techniques fail to take semantic relationships into account. Uncommon relevant terms, not containing the input query term, are often ignored. Techniques based purely on query-logs fail to explore new words, not very frequently correlated by query log data. To address the aforementioned problems, we suggest a new technique called TermsNet. TermsNet grasps the underlying semantics among words and suggests new keywords from the terms corpus. It does not require any query log data. This technique can easily adapt to trends. Newer terms can be simply added to the existing graph and made available for querying and suggestion, irrespective of whether that term has become popular among users (and hence query logs) or not. Even uncommon terms show up in the results if they are relevant. The technique scales very well with data, and results improve with more input words.

3. TermsNet

This section introduces TermsNet, our approach to keyword research. TermsNet leverages search engines to determine relevance between terms and captures their semantic relationships as a directed graph. By observing neighbors of a term in such a graph, we generate the common as well as the nonobvious keywords related to a term.

Terms are too short by themselves to be examined for similarity using the traditional document similarity measures like cosine coefficient. Similarity needs to be measured by examining the context of the two terms as proposed in [6]. [1] and [3] suggest that a query can be effectively expanded by augmenting it with additional terms based on documents retrieved from a search on that query, or by using an available thesaurus. We adopt the idea of using search engine to expand context of queries. However, unlike [6], the problem here is keyword research and not query expansion/suggestion. The query suggestion system in [6] is used to suggest on an average 2-3 words per query. Keyword research needs hundreds of keyword suggestions to be effective. TermsNet provides a framework suitable to explore a large number of similarity relationships simultaneously, while giving a large number of suggestions for queries. In [6], the system maintains a list of precomputed suggestions based on similarity kernel function and dynamic addition of new words is difficult. In keyword research, if a new word enters the corpus, it is important to be able to find it immediately for suggestions. TermsNet provides a framework in which new words can be dynamically added and suggested with very little effort.

We represent each term using a characteristic document containing text-snippets from top 50 search-hits for that term. A text snippet is the sentence containing the searched term. In cases where the sentence is too short or too long, we consider words before and after the term, keeping length of the snippet constant. Thus, the characteristic document is considered as a representation of the term and its context.

Here, we introduce the notion of directed relevance. The relevance relationship between terms is directed, i.e., term A may strongly suggest term B, but not vice versa. For example the term 'eurail' strongly suggests 'europe' and 'railways', but the term 'europe' or 'railways' may not suggest 'eurail' with the same strength. Here we say that 'eurail' is a highly relevant, yet nonobvious suggestion for both 'europe' and 'railways' but not vice versa. To enforce this notion of directed relevance, we redefine short-text similarity. Instead of considering the degree of overlap between the characteristic documents of terms, we measure relevance of B to A as the frequency of term B observed in the characteristic document of term A and vice versa. A high frequency of term B in term A's characteristic document tells us that term A suggests term B. Since frequency of A in B's characteristic document may be different from the frequency of B in A's document, the relevance of A to B is different from the relevance of B to A, thus inducing the notion of directed relevance.

Using the characteristic document, we calculate the directed relevance between terms and express the result as a directed graph. In this graph, nodes represent terms. A directed edge from node A to node B, represents a relevance relationship, i.e., word A suggests word B. Edge weights reflect the relevance values calculated using the characteristic document. The outgoing edges from a node represent the terms suggested by source node, while incoming links represent terms that suggest the node. To generate keywords related to a given input query term we follow outgoing links as well as trace back the incoming links to get relevant terms. The outgoing links give terms that are suggested by source term. These are terms that can also be extracted by simply looking for the high frequency terms in the characteristic document. Most of these terms, though relevant, are not very good suggestions. This is because they do not suggest the source word, but are rather suggested by the source. Nevertheless, concatenated with the source word, they are useful to generate a new phrase keyword.

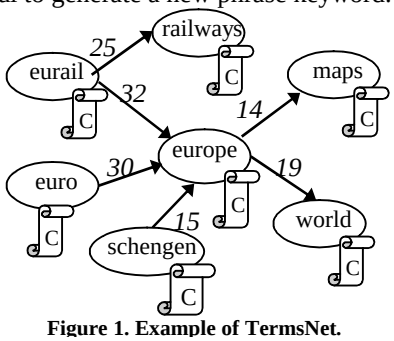

Figure 1. Example of TermsNet.

The more interesting terms are those from incoming edges. The incoming edges to a term represent the nonobvious, yet strong relevance relationships. In Association Rule Mining terms, an edge E:A→B with weight 'x' represents the confidence of relationship A→B. In such a case, suggesting 'A' for query 'B' adds value in terms of both, relevance and nonobviousness. Note that the confidence of B→A may be altogether different. The candidate terms for suggestion are ranked using their edge weights. Results are further tweaked using the concept of Inverse Document Frequency (IDF), i.e., down-weighting a term that is too general. In this case general terms are ones that suggest too many other terms, i.e., have a high outdegree. Hence, edge weights are down-weighted, by a factor of logarithm of number of outgoing edges using from that node. We use logarithm to smooth out the excessive effect of IDF for very large outdegree. Thus, the quality of suggestion 'x' for query 'q' is

$$Q(x, q) = w_{x,q} / (1+\log(1+\sum w_{x,i}))$$ where each i is an outneighbor of 'x'

Suggested keywords from TermsNet are ranked using this quality measure.

4. Experiments

In this section, we present experimental setup and results. TermsNet is implemented in Java. For generating the characteristic document for each term, we use Google Search APIs. For TermsNet, we reinforce term-term relationships on top of an underlying inverted index framework for term-document relationships. We use Apache Lucene Library for implementing indexing and querying support over the TermsNet. For these experiments, we primarily focus on queries related to three broad topics popular among advertisers, viz., travel, car-rentals, and mortgage. We ran experiments on an input set of 8,000 search terms, picked randomly from webpages relevant on the three broad targeted topics. These terms can be picked from query logs, if available. Keyword suggestion results are obtained for 100 benchmark queries based on the three broad targeted topics. For comparative evaluation, we select tools based on the different techniques used for generating words. We use AdWords Specific Word Matches [2], AdWords Additional Keywords [2], Overture Keyword Selection Tool [4], Meta-Tag Spider [5] and Related-Keywords list from Metacrawler [5]. The underlying techniques for these are discussed in detail in Section 2. Since the objective here is to compare techniques and not the tools, AdWords Specific Word Matches and AdWords Additional Keywords are treated as separate entities.

The benchmark queries were run on TermsNet and other tools. Table 1 shows top results for a sample query (viz. flights), from the various tool used for comparison. As indicated by the output, TermsNet tends to generate highly relevant, yet nonobvious terms consistently.

Table 1: Results for a sample query, viz., 'flights'

AdWords Additional	AdWords Specific	Meta-Tag Spider	Meta-Crawler	Overture	TermsNet
Airfare	Flights	real time flight	air travel	flight	cheap flights
airfares	cheap flights	arrivals	airline discount	cheap flight	airline flights
airlines	airline flights	airfare	tickets	las vegas flight	air newzealand
cyprus	cheap airline flights	flights	airline fares	flight tracker	flight prices
fights	cheap international flights	flight	airline tickets	flight to orlando	bmibaby
holidays	flights to europe	cruises	airline tickets	flight to london	globespan
trains	business class flights	us flight arrivals	under 100	flight to new york	low cost airlines
aer	flights new york	flight arrivals	american airlines	airline flight	united airlines
aeroflot	australia flights	state map	bargain flights	flight to los angeles	airline-consolidators
aeromexico	cheap flights to europe	flight arrival	bmibaby	flight 93	charter flights
aircanada	cheap flights to orlando	map	british airways	flight to fort	airfare
alicante	cheap flights las vegas	flight cancellations	british airways	lauderdale	flight reservations
bwia	track flights	arrival times	flights	light of the phoenix	cathay pacific
flys	flights florida	arrival delays	british airways	flight to honolulu	british midland
goa	flights europe	delays	home page	flight to chicago	airways
heathrow	las flights	flight departure	british airways	flight to miami	discount airfare
icelandair	cheap flights to australia	vacation packages	timetable		flight tickets
bookings		street map	british midland		jet2
consolidator			budget airline		travelocity

5. Evaluation and Results

Each keyword suggestion was given two ratings, viz., Relevance and Nonobviousness. A relevance rating of Relevant/Irrelevant was provided by 5 human evaluators, who are graduate students at Stanford and familiar with the requirements of this technique. For nonobviousness rating, we define nonobvious term as a term not containing the seed keyword or its variants sharing a common stem. Using the standard Porter Stemmer [8], we marked off nonobvious words, without involving human evaluators. Each technique was evaluated using the measures of average precision, average recall, and average nonobviousness.

Average Precision measures the goodness of a technique in terms of the fraction of relevant results returned. It is defined as the ratio of number of relevant keywords retrieved to number of keywords retrieved. Average Recall is the proportion of relevant keywords that are retrieved, out of all relevant keywords available. The problem with determining exact recall is that the total number of relevant keywords is unknown. Hence we approximate this as the size of the union of relevant results from all techniques. Due to this approximation, recall values, though imperfect in the absolute sense, are useful to compare techniques. We define average nonobviousness as the proportion of nonobvious words, out of retrieved relevant words. Precision, Recall and Nonobviousness are calculated for each query and their respective results are averaged over each technique.

Harmonic mean of precision and recall is the traditional F-measure. Since, we aim to maximize precision P, recall R and nonobviousness N, we define four new F-measures, viz., F(PR), F(PN), F(RN) and F(PRN). These act as a measurable value of overall goodness of a technique. Among the compared techniques, TermsNet achieved the highest values for all four F-measures.

Table 2 presents the measures calculated for every technique. As is indicated by the values, TermsNet outperforms other techniques on most of the measures. More importantly, TermsNet ranks highest for all the four F-measures. Higher ranking in F-measure indicates that the technique manages to achieve high overall scores in all the considered factors.

Since AdWords Specific Matches and Overture Keyword Selection Tool output only queries containing the seed term, almost all the suggested words are relevant, but too obvious. Hence, these techniques are skewed towards good precision, but poor nonobviousness. Meta-tags from a webpage may or may not contain highly relevant terms. This technique does well with the recall, but underperforms at precision and nonobviousness. Results from the Metacrawler's related keyword list are usually highly relevant, thus getting high precision value and reasonably high nonobvious too. However, MetaCrawler tends to give much fewer results and hence low recall. The keyword research problem requires large number of keyword suggestions to be practically useful. Hence, the MetaCrawler can, at most, be useful as a preprocessing stage to other techniques with higher recall. TermsNet captures relevance very well because the underlying graph is built using semantic relationships and co-occurrence. It has a relatively high recall, as it tends to give a fair chance to all terms in the underlying graph. Larger the underlying graph, greater the recall. Nonobviousness too is correctly captured because it exploits the incoming links to a term. Thus, TermsNet does better than other techniques on all the F-measures.

In another experiment, we measured the quality of

Table 2: Evaluation of different techniques.

	Ad-Broad	Ad-Spec	Meta-Tags	MetaCrawler	Overture	TermsNet
Avg. Precision	0.636364	**1**	0.479675	0.94	**1**	0.788043
Avg. Recall	0.196	0.254	0.118	0.094	0.201	**0.58**
Avg. Nonobv. Rating	**1**	0	0.559322	0.744681	0	0.913793
F (PR)	0.149847	0.202552	0.094703	0.085455	0.16736	**0.334101**
F (PN)	0.388889	0	0.258223	0.415509	0	**0.423136**
F (RN)	0.16388	0	0.097443	0.083464	0	**0.354801**
F (PRN)	0.068069	0	0.027363	0.036994	0	**0.183038**

suggestions over different intervals of ranked results. In particular, we observed how average precision and average nonobviousness vary with increasing number of ranked keywords returned. These two measures reflect the confidence of the ranked results in different ranges.

Figure 2 shows results of these experiments. The average nonobviousness rating is almost constant at about 0.9 over all the intervals. This indicates that the technique consistently maintains nonobviousness in its output. Since nonobviousness is almost constant, we use precision to determine a good cutoff threshold.

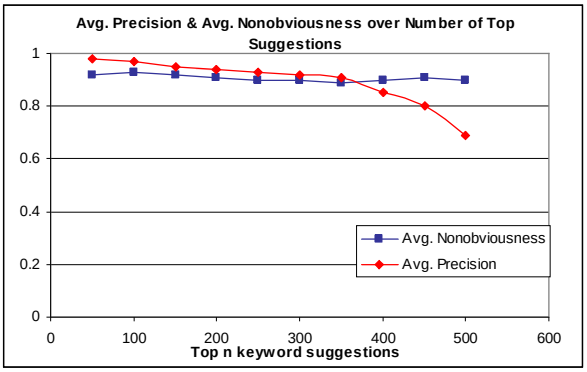

Figure 2: Quality of keywords over different ranked intervals

The average precision is very close to 1.0 for the top 50 suggestions. The curve gradually slopes downwards as we include lower ranked results. After about rank 350, average precision begins to fall more rapidly. As we see more and more lower-ranked suggestions, irrelevant keywords increase in number, thereby lowering the average precision. In other words, we see more false positives in our output with increasing number of suggestions. Since the requirement of keyword research is to get very large number of keywords, we can tolerate a fair number of false positives. However, a reasonable cutoff threshold would be about 0.75, after which the statistical relevance relationship becomes too weak.

The initial gentle slope in average precision indicates that the keyword ranking done by TermsNet is quite stable on relevance till a large number of keywords. As the number of terms in the TermsNet increases, we have more number of keywords above the precision cutoff threshold. This initial almost flat nature of precision curve indicates that even with larger number of terms, a good ranking like this one can ensure consistently high relevance among the top ranked keywords.

6. Conclusion

We have demonstrated the effectiveness of TermsNet in generating nonobvious and relevant keywords for keyword research. Possible extensions to TermsNet include incorporating keyword frequency into the current setting by storing term frequency information of TermsNet and using it as a factor in the ranking of suggested terms. Iteration can be added to TermsNet to improve recall or to explore distant relations among terms.

TermsNet can be effectively applied to problems like finding related movies, academic papers, people, etc. It can be used for automatic thesaurus generation when input terms are dictionary words or for organizing pictures based on their tags. Our broader objective is to be able to extract related items by identifying their associations using the World Wide Web. Thus, TermsNet has good potential to be an effective technique for problems where textual relationships between objects need to be extracted.

7. References

[1] C. Buckley, G. Salton, J. Allan, A. Singhal, “Automatic query expansion using SMART:TREC3”, TREC 3, 1994, 69-80

[2] Google AdWords Keyword Tool, adwords.google.com/select/KeywordSandbox

[3] Mitra, M., Singhal, A., Buckley, C., “Improving Automatic Query Expansion”, SIGIR1998, 206-214.

[4] Overture Keyword Selection Tool, inventory.overture.com/d/searchinventory/suggestion

[5] Rapid Keywords Tool, http://www.rapidkeyword.com

[6] M. Sahami, T. Heilman, “A Web-based Kernel Function for Matching Short Text Snippets”, Workshop on "Learning in Web Search", ICML 2005

[7] WordTracker Tool, www.wordtracker.com

[8] Porter, M.F. Porter, “An algorithm for suffix stripping”, Program, 14(3):130-137, July 1980