Context Sensitive Synonym Discovery for Web Search Queries

Xing Wei Fuchun Peng Huishin Tseng Yumao Lu Benoit Dumoulin
Yahoo! Labs
701 First Avenue, Sunnyvale, California, USA, 94089
xwei,fuchun,huihui,yumaol,benoitd}@yahoo-inc.com

ABSTRACT

We propose a simple yet effective approach to context sensitive synonym discovery for Web search queries based on co-click analysis; i.e., analyzing queries leading to clicking same documents. In addition to deriving word based synonyms, we also derive concept based synonyms with the help of query segmentation. Evaluation results show that this approach dramatically outperforms the thesaurus based synonym replacement method in keeping search intent, from accuracy of 40% to above 80%.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: Synonym discovery, query reformulation

1. INTRODUCTION

Synonyms are words or expressions of the same language that have the same or nearly the same meaning in some or all senses [1]. Automatically discovering synonyms from text has been active topics in a variety of language processing tasks [3, 7, 1, 2, 4]. Most existing work is to create a general purpose synonym thesaurus without targeted applications [5]. However, it is unclear how general synonyms can help a particular application. In the context of Web search, we want to find synonyms that can express the same search intent of users. Discovering synonyms for Web search have at least the following challenges:

1. synonym discovery is context sensitive. Although there are quite a few manually built thesauri available to provide high quality synonyms [2], most of these synonyms have the same or nearly the same meaning only in some senses. If we simply replace them in search queries, it is very easy to trigger search intent drift, a very bad search experience for users. For example, "baby" and "infant" are treated as synonyms in many thesauri, but "Santa Baby" has nothing to do with "infant" not only it is a song title, which is an entity that needs special handling, but also the meaning of "baby" in this entity is different than the usual meaning of "infant".

2. context can not only limit the use of synonyms as above, but also broaden the traditional definition of synonyms. For instance, "dress" and "attire" sometimes have nearly the same meaning, even though they are not associated with the same entry in many thesaurus; "free" and "download" are far from synonyms in traditional definition, but "free cd rewriter" may carry the same query intent as "download cd rewriter".

3. there are many new synonyms developed from the Web over time. "Playstation 3" and "ps3" were not synonyms twenty years ago; "snp newspaper" and "snp online" carry the same query intent only after snponline.com was published. Thus a static synonym list is less desirable.

In summary, synonym discovery for Web search is different from traditional thesaurus mining; it needs to be context sensitive to keep the same search intent, and is time sensitive so we need to update the dictionary timely. To address these problems, we conduct context based synonym discovery from co-clicked queries, i.e., queries that share similar document click distribution. The intuition of discovering synonyms from co-clicked data is that queries with similar clicked URLs tend to be related and carry similar intents, which are often formulated with synonyms. Synonyms discovered from co-clicked documents is context sensitive; as we aggregate over many queries, the distribution of clicked documents reflects pretty well the search intent. We can obtain search queries on a daily basis, thus the synonyms mined from search queries reflects most recent synonyms.

2. CO-CLICKED QUERY CLUSTERING

Clustering has been extensively studied in many applications, including query clustering [8]. One of the most successful techniques for clustering is based on distributional clustering [3, 4]. We adopt a similar approach to our co-clicked query clustering. Each query is associated with a set of clicked documents, associated with the number of views and clicks. We then compute the distance between a pair of queries by calculating the Jensen-Shannon(JS) divergence between their clicked URL distributions. We start with that

[1]according to the definition by www.merriam-webster.com

CIKM'09, November 2–6, 2009, Hong Kong, China.

every query is a separate cluster, and merge clusters greedily. After clusters are generated, pairs of queries within the same cluster can be considered as co-clicked/related queries with a similarity score computed from their JS divergence.

3. SYNONYM DISCOVERY BASED ON CO-CLICKED QUERY

Synonyms discovered from co-clicked queries have two aspects of word meaning: (1) general meaning in the language (2) specific meaning in the query. These two aspects are related. For example, if the two words are more likely to carry the same meaning in general, then they are more likely to carry the same meaning in specific queries; on the contrary, if two words often carry the same meaning in a variety of specific queries, then we tend to believe that the two words are synonyms in general language.

We develop our model based on the above two aspects. First, we try to get the general meaning of two words from their meanings in all specific queries. As we described above, the assumption here is that if the two words carry the same meaning in many specific queries, then they are likely to be synonyms for aspect (1). We consider all the co-clicked queries with the word and sum over, as in Eq. 1

$$\frac{\sum_k sim_k(w_i \rightarrow w_j)}{\sum_{w_j} \sum_k sim(w_i \rightarrow w_j)} \tag{1}$$

where $sim_k(w_i \rightarrow w_j)$ represents the similarity score of a query q_k that aligns w_i to w_j. So intuitively, we aggregate scores all query pairs that align w_i to w_j, and normalize it to a probability ($P(w_j|w_i)$) over the vocabulary.

Then, the query is taken into consideration to catch the specific meaning in aspect (2). We define the probability of reformulating w_i with w_j for query q_k to be the similarity score as in Eq. 2

$$P(w_j|w_i, q_k) = sim_k(w_i \rightarrow w_j) \tag{2}$$

Last, we combine the above two steps. We have two sets of estimations for the synonym probability, which is to reformulate w_i with w_j. One set of values are based on general language information and another set of values are based on specific queries. We applied linear combination in log scale to combine the two probabilities as in Eq. 3

$$\begin{aligned}\log P_{q_k}(w_j|w_i) =& \lambda * \log P(w_j|w_i) + \\ & (1-\lambda) * \log P(w_j|w_i, q_k)\end{aligned} \tag{3}$$

The simple word alignment strategy we used can only get the synonym mapping from single term to single term. But there are a lot of phrase-to-phrase, term-to-phrase, or phrase-to-term synonym mappings in language, such as "babe in arms" and "infant", "nyc" and "new york city". We perform query segmentation on queries to identify concepts from queries based on an unsupervised segmentation model [6]. Query segmentation not only gives concept based alignment, but also can improve the precision of alignment. For example, "baby clothing stores" will not be aligned with "baby favorite stores" after segmentation even they contain the same number of words.

4. EXPERIMENTS

4.1 Evaluation Metrics

Because we are more interested in the application of reformulating Web search queries, our guideline to the editorial judgment focus on the query intent change and context-based synonyms. For example, "transporters" and "movers" are good synonyms in the context of "boat" because "boat transporters" and "boat movers" keeps the same search intent, but "ocean" is not a good synonyms to "sea" in the query of "sea boss boats" because "sea boss" is a brand name and "ocean boss" does not refer to the same brand. Results are measured with accuracy by number of discovered synonyms (which reflects coverage). And accuracy are defined as

$$Accuracy = \frac{n_{correct}}{N} \tag{4}$$

where $n_{correct}$ is the number of correctly discovered synonyms, and N is the number of all discovered synonyms.

We target on higher accuracy and larger coverage, but generally, discovering more synonyms would lead to more errors, which means lower accuracy.

4.2 Data

A period of Web search query log with clicked URLs are used to generate co-clicked query set. After word alignment, which extract the co-clicked query pairs with same length and only one different unit, we have around 12.1M unsegmented query pairs and 11.9M segmented query pairs.

We randomly sampled 42K queries from two weeks of query log, and evaluate the effectiveness of our synonym discovery model with these queries. To test the synonym discovery model built on the segmented data, we segmented the queries first before sending the data as the evaluation set.

4.3 Synonym Discovery Accuracy

In this section we present WordNet thesaurus based query synonym discovery, co-clicked based term-to-term query synonym discovery, and co-click concept based query synonym discovery.

4.3.1 Thesaurus-based synonym replacement

The WordNet thesaurus-based synonym replacement is a baseline of our approach. For any word that has synonyms in the thesaurus, thesaurus-based synonym replacement will rewrite the word with synonyms from the thesaurus.

Thesaurus-based synonym replacement suffers from missing of context. Although thesaurus can provide clean information, it has only knowledge for single words. The context plays an important role in synonym discovery, and thesaurus-based synonym replacement without considering context often brings too much errors and noise. Our experiments show that only less than **46%** of the discovered synonyms are correct synonyms in query. Although 27k synonyms were discovered from the test set, which are much more than the number of synonyms our approach discovered (see Section 4.3.2 and Section 4.3.3), the accuracy is too low to be used for Web search queries.

4.3.2 Co-clicked query-based context synonym discovery

Figure 1 demonstrates how accuracy changes with the number of synonyms. Y axis represents the percentage of

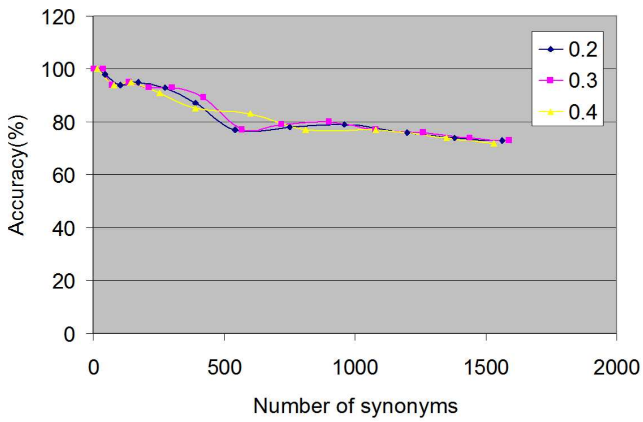

Figure 1: Accuracy versus number of synonyms. Mixture weight λ=0.2, 0.3, or 0.4.

correctly discovered synonyms; X axis represents the number of discovered synonyms, including both of correct ones and wrong ones. The three different lines represents three different parameter settings of mixture weights(λ in Eq. 3, which is 0.2, 0.3, or 0.4 in the figure) selected by experience. The figure shows accuracy drops by raising the number of synonyms. More synonym pairs tend to imply lower accuracy. From Figure 1 we can see:

1. the effectiveness of synonym discovery is not very sensitive to the mixture weight(if set in a reasonable range);

2. the effectiveness of synonym discovery is sensitive to the threshold, which leads to different numbers of discovered synonyms. By loosening the threshold to get more synonyms, the accuracy decreases from **100%** to less than **80%** (we are not interested in accuracies lower than 80% due to the high precision request of Web search task, so the graph contains only high-accuracy results). This trend also confirms the effectiveness of our approach since the threshold is a significant factor in synonym discovery and the accuracy increases by tighting the threshold.

4.3.3 *Concept based context synonym discovery*

We present results from our model based on segmented co-clicked query data in this section. The modeling part is the same as the one for Section 4.3.2, and the only difference is that the data were segmented.

Figure 2 shows the accuracy of synonyms by number of discovered synonyms. As in Section 4.3.2, by applying different thresholds as cut-off lines to Eq. 3, we get different numbers of synonyms from the same test set, and a looser threshold gives us more synonym pairs with lower accuracy. Y axis in Figure 2 represents the percentage of correctly discovered synonyms and X axis represents the number of discovered synonyms. We have shown in Section 4.3.2 that the mixture weight is not an influential factor within reasonable range, so we present only the result with one mixture weight in Figure 2.

Same as in Section 4.3.2, the figure shows that the accuracy of synonym discovery is sensitive to the threshold. Loosing the threshold to get more synonyms decreases the accuracy. Again, it confirms that our model is effective and setting threshold to Eq. 3 is a feasible and sound way to discover not only single term synonyms but also phrase synonyms.

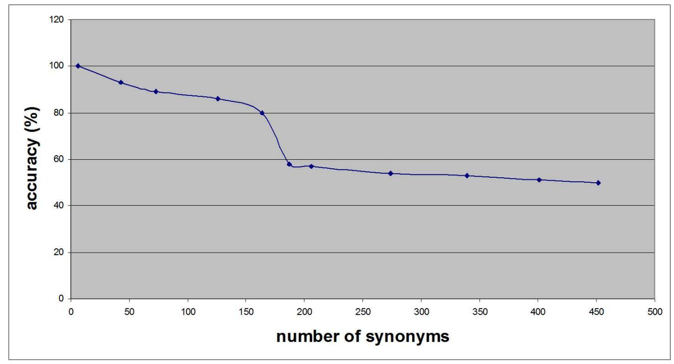

Figure 2: Accuracy versus number of synonyms. Mixture weight λ=0.3.

Table 1 shows representative examples of query synonyms with the thesaurus-based synonym replacement, context sensitive synonym discovery, and concept based context sensitive synonym discovery. The upper part of each section shows positive examples (query intents remain the same after synonym replacement) and the lower part shows negative examples (query intents change after synonym replacement).

5. DISCUSSION AND ERROR ANALYSIS

From Table 1, we can see that our model can catch not only traditional synonyms, which are the synonyms that can be found in manually-built thesaurus, but also context-based synonyms, which may not be treated as synonyms in dictionaries or thesaurus.

However, the click data themselves contain huge amount of noise. Although they can reflect the users' intents in big picture, in many specific cases synonyms discovered from co-clicked data are biased by the click noise. In our application – Web search query reformulation with synonyms, accuracy is the most important thing and thus we are interested in error analysis. The errors that our model made in synonym discovery are mainly caused by the following reasons:

1. *popular concepts:* There are some concepts that are well accepted such as "cnn" means "news" and "amtrak" means "train". And users searching for "news" tend to click CNN web site; users searching for "train" tend to click Amtrak web site. With our model, "cnn" and "news", "amtrak" and "train" are discovered to be synonyms, but this may hurt the search of "news" or "train" in general meaning.

2. *same clicks by different intents:* Different intents/meanings could results in same or similar clicks. Query "antique style wedding rings" and "antique style engagement rings" carry different intents, but very usually, these two different intents can lead to the clicks on the same store web sites. Other examples include "booster seats" and "car seats", "brighton handbags" and "brighton shoes". For these examples, clicks on Web URLs are not precise enough to reflect the detailed difference of language concepts.

3. *dominant user intents:* Most people searching for "airline travel restrictions" are looking for "airline baggage restrictions". So these two queries have similar clicked-URLs. But "travel" and "baggage" are not synonyms in language.

Original Query	New Query with Synonyms	Intent
Examples of thesaurus-based based synonym replacement.		
basement window wells drainage	basement window wells drain	same
billabong boardshorts sale	billabong boardshorts sales event	
bigger stronger faster documentary	larger stronger faster documentary	
colored contacts	coloured contacts	
how do u tighten the shift bands on a auto transmission	how do u tighten the shift bands on a car transmission	
paxil	paroxetime	
yahoo	hayseed	different
maryland judiciary case search	maryland judiciary pillowcase search	
free cell phone number lookup	free cell earpiece number lookup	
aim mail	purpose mail	
free texas quick claim form	free texas quick claim organise	
win star casinos	win champion casinos	
Examples of term-to-term synonym discovery.		
airlines jobs	airlines careers	same
area code finder	area code search	
acai berry	acai fruit	
business definitions	business terms	
countrywide loans	countrywide mortgage	
cox webmail	cox email	
acai berry	acai juice	different
ace	hardware	
crest toothpaste coupon	crest whitestrips coupon	
delta faucet repair	delta faucet parts	
dish	dishnetwork	
dell laptops	dell computers	
Examples of concept based synonym discovery.		
ae	american_eagle outfitters	same
apartments_for_rent	apartment_rentals	
arizona time_zone	arizona time	
bank of america online_banking	bank of america online	
brown recluse spider_bite	brown recluse bite	
crossword_puzzles	crossword	
cortrust bank credit_card	cortrust bank mastercard	different
david_beckham	beckham	
dodge_caliber	dodge	
apartments_for_rent	apartment	
brown recluse spider bite pictures	spider bite pictures	
california health_department	california medicaid	

Table 1: Examples of query synonym discovery

In these cases, popular user intents dominate and biased the meaning of language, which cause problems.

4. *antonyms:* Many context-based synonym discovery methods suffer from the antonym problem, because antonyms can have very similar contexts. In our model, the problem has been reduced by integrating clicked-URLs. But still, there are some examples, such as "spyware" and "antispyware", resulting in similar clicks. To learn how to "protect a web site", the user will often need to learn what are the main methods to "attack a web site", these different-intent pairs lead to same clicks because different intents do not have to mean different interests in specific cases.

For future work, we are investigating using these synonyms in improving search relevance. Our preliminary results show this is promising.

6. REFERENCES

[1] M. Baroni and S. Bisi. Using Cooccurrence Statistics and the Web to Discover Synonyms in a Technical Language. In *LREC*, 2004.

[2] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Mass., 1998.

[3] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL-98*, pages 768–774, 1998.

[4] F. Pereira, N. Tishby, and L. Lee. Distributional Clustering of English Words. In *Proceedings of ACL-93*, pages 183 – 190, 1993.

[5] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of COLING/ACL-06*, pages 801–808, 2006.

[6] B. Tan and F. Peng. Unsupervised Query Segmentation using Generative Language Models and Wikipedia. In *Proceedings of the 17th International World Wide Web Conference (WWW)*, pages 347–356, 2008.

[7] P. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning*, 2001.

[8] J.R. Wen, J.Y. Nie, and H.J. Zhang. Query Clustering Using User Logs. *ACM Transactions on Information Systems*, 20(1):59–81, 2002.