Which "Apple" Are You Talking About ?

Mandar A. Rahurkar
University of Illinois
rahurkar@uiuc.edu

Dan Roth
University of Illinois
danr@cs.uiuc.edu

Thomas S. Huang
University of Illinois
huang@ifp.uiuc.edu

ABSTRACT

In a higher level task such as clustering of web results or word sense disambiguation, knowledge of all possible distinct concepts in which an ambiguous word can be expressed would be advantageous, for instance in determining the number of clusters in case of clustering web search results. We propose an algorithm to generate such a ranked list of distinct concepts associated with an ambiguous word. Concepts which are popular in terms of usage are ranked higher. We evaluate the coverage of the concepts inferred from our algorithm on the results retrieved by querying the ambiguous word using a major search engine and show a coverage of 85% for top 30 documents averaged over all keywords.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Clustering **General Terms:** Algorithms, Experimentation.

Keywords

Wikipedia, Concepts, Clustering

1. INTRODUCTION AND PRIOR WORK

Consider an ambiguous query [3] "apple" queried on a search engine. The search relevance can be improved if we know which "apple" user is interested in. One of the better ways to present the relevant search results would be to cluster the results wherein we have a cluster for each distinct concept in which the word apple can be used, e.g, *Apple (corporation)* or *Apple (fruit)*. However most clustering algorithms require the number of clusters as an input. This requires multiple iterations seeking an optimal number based on some statistical criteria which may or may not guarantee consistency within and across different clusters. Thus, in this example as well as other applications including word sense disambiguation, knowing number of unique concepts in which a word can be used would be an invaluable asset. In this poster we propose an algorithm to determine all possible unique concepts for a given ambiguous word. We do so by using the Wikipedia. Wikipedia has been recently used for named entity disambiguation [1] and word sense disambiguation(WSD) [2] tasks. In [2] contextual text surrounding the ambiguous word in addition to the word itself is used without explicitly enumerating the distinct concepts of the ambiguous word. In case of web queries which are typically very short such an approach may not be an option.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. DATA

Wikipedia is a collection of articles where each article defines and describes an entity or an event. Each article may have several hyperlinks to the other pages within or outside Wikipedia and is uniquely referenced by a title. Title is composed of one or more words and occasionally an explanation in parenthesis clarifying the context of the article. For example, the article for mercury with the meaning of "automobile" has the unique identifier *mercury (automobile)*. Ambiguous surface form is hyper linked to the appropriate article using a pipe, e.g., link from the word *orange* to an article on *Orange Color* (if applicable), as in [[Orange (Color)|Orange]]. This can generally be represented as [[Concept Identifier| Surface form]] pair. The phrase "concept identifier" and the word concept is used interchangeably. We use this structural information within the Wikipedia to identify possible concepts for a given ambiguous word. Since these links have been manually created and reviewed by a large diverse audience, they are accurate in referencing the article clarifying the context in which the surface form has been used.

3. ALGORITHM

We parse the Wikipedia[1] to extract all the hyperlinked occurrences of a word. While parsing, the disambiguation pages, pages associated with dates and pages enumerating the lists are excluded. Wikipedia, being updated regularly by a diverse group of people reflects the realistic use of these concepts. Therefore, we use the occurrence count in the first pass to obtain the ranked list $\mathcal{R}'$ of concepts associated with a keyword such that $\mathcal{R}' = \{C_1, C_2, \ldots, C_n\}$ and rank of $C_i >$ rank of C_j for $i < j$. Top 20 such concepts for ambiguous keyword "Bush" are enumerated in table 1.

Most surface forms are found to associate with a large number of concepts, e.g., surface form *Orange* is associated with about 150 concepts, *Mercury* with 110 concepts etc. However large number of these concepts might not be unique as seen in the table 1 where concepts "george w. bush", "george walker bush" and "george w bush" represent the same concept. We re-rank $\mathcal{R}'$ filtering it for the duplicate concepts.

3.1 Re-Ranking the concept list

Consider a directed graph $G = \{V, E\}$ where each vertex V, represents a Wikipedia article or a concept. There exists

[1]XML file dump as on September 07

Table 1: Top 20 concepts for ambiguous query "bush" before the list is filtered and re-ranked.

Rank 1-10	*Rank 11-20*
george w. bush	bush family
bush (band)	bush alaska
george h. w. bush	george walker bush
the bush	alaskan bush
shrub	outback
george h.w. bush	uss bush (dd-529)
bush (canadian band)	bush, alabama
forest	bush plane
george herbert walker bush	woody plant
george w bush	bush, illinois

an outgoing edge from two vertices V_i and V_j, $V_i \rightarrow V_j$, if there is an outgoing hyperlink from article i to j in the Wikipedia. Example graph is shown in figure 1, where C_1 and C_2 might represent "george w. bush" and "george walker bush". Given a ranked list $\mathcal{R}'$ containing concepts $C_1, C_2, \ldots, C_k$ we seek filtering measures to re-rank the list and prune the duplicate concepts. If concept $C_i \subseteq C_j$ under some measure M, then the concept C_i is subsumed by the concept C_j. If a concept C_j subsumes concepts $C_i, C_k, \ldots, C_l$, the counts of these concepts are added to C_j thereby increasing its score and possibly the rank. Ranked list $\mathcal{R}'$ is processed sequentially starting with the concept node at the top of $\mathcal{R}'$. Graph beginning from this node is parsed in a depth first search fashion to identify concepts C_i, where $i > j$ such that $C_i \subseteq C_j$ under the measure M defined later. Since there are cycles in the graph we halt processing of a node if we encounter its ancestor. We now compute three measures and combine them to re-rank $\mathcal{R}'$. The first measure checks for the existence of a bi-directional link between the two concepts.

$$M_b(C_i, C_j) = \mathbf{I}((C_j \rightarrow C_i) \wedge (C_i \rightarrow C_j)) \quad (1)$$

where, $\mathbf{I}$ is an Indicator variable. Thus, a concept C_j may subsume C_i if $M_b = 1$. M_b measure however is a greedy measure and concept C_i may subsume weakly related concept C_j due to a presence of a bi-directional link, e.g., *Orange Color* and *Syracuse Orange*.

If two concepts share a subset of outbound and inbound links they are likely to be similar. The second and third measure (not shown) count the overlap between the inbound and outbound links. Operators $In(C)$ and $Out(C)$ return a set of inbound and outbound links for concept C and $|x|$ returns the cardinality of the set x. M_{OB} is similar to M_{IB} but defined instead on the outbound links.

$$M_{IB}(C_m, C_n) = max\left(\frac{|In(C_m) \cap In(C_n)|}{|In(C_m)|}, \frac{|In(C_m) \cap In(C_n)|}{|In(C_n)|}\right)$$

The fourth measure combines these three measures,

$$M(C_m, C_n) = M_b(C_m, C_n) \times \left(\alpha_1 M_{IB}(C_m, C_n) + \alpha_2 M_{OB}(C_m, C_n)\right)$$

where, $\alpha_1 + \alpha_2 = 1$.

Measure M yields values in $[0, 1]$, and a higher value indicates more similarity. A concept in list $\mathcal{R}'$, C_j, subsumes C_i, where $i > j$ if $M(C_m, C_n) > PVal$. $PVal$ is determined empirically to be 0.24. Concepts which occur less than 5 times are discarded. The pruned concept lists for the ambiguous words *bush* and *mercury* are shown in table 2.

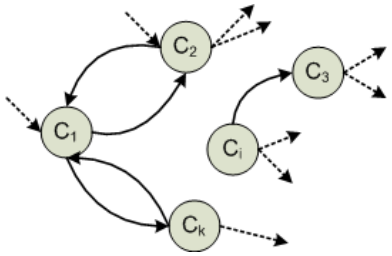

Figure 1: Concept Graph

Table 2: Concepts after filtering the list $\mathcal{R}'$ for keywords bush and Mercury

Bush	*Mercury*
George W. Bush	Mercury (element)
Bush (band)	Mercury (planet)
Bush alaska	Mercury (records)
Forest	Mercury (mythology)
Bush LA	Mercury (automobile)

4. RESULT AND CONCLUSION

We evaluate the ranked list of concepts $\mathcal{R}$ by examining its coverage over top 30 results returned by a major search engine on querying the ambiguous words from [2, 3]. Human annotators were shown the retrieved pages and asked to assign the concept to a page from the list $\mathcal{R}$ which best describes the content on that page, e.g., *Orange telecom* for a page related to "Orange mobile company". In addition to concepts in the list $\mathcal{R}$ associated with the ambiguous word, annotator could also assign the labels "Can't Say", "Other: Not defined here" and "Tech Error" in cases when they could not reliably identify a concept, or if the concept on webpage was not mentioned in the list or if an error occurred in loading a webpage. We obtained a percentage coverage of over **85%** for top 30 documents and **89%** for top 5 documents averaged across all the keywords. Coverage is low for the concept *jordan* because jordan in Wikipedia has not been used to refer to "michael jordan, professor" or several other firms by that name which were part of the retrieved results. Coverage should improve to an extent on using the latest version of the Wikipedia dump.

5. REFERENCES

[1] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL 2007*, pages 708–716, 2007.

[2] R. Mihalcea. Using wikipedia for automatic word sense disambiguation. *NAACL*, 2007.

[3] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. *SIGIR*, July 2004.

Table 3: Percentage coverage for top 30 results obtained on issuing an ambiguous query.

Query	*Avg. Coverage*	*Query*	*Avg. Coverage*
apple	100.00	lincoln	73.33
bar	83.33	matrix	66.67
bush	90.00	orange	80.00
clinton	73.33	quotes	90.00
ford	100.00	saturn	96.67
jaguar	86.67	tiger	86.67
jobs	96.67	ups	93.33
jordan	66.67		