Generating Hypotheses from the Web

Wei Jin
Dept. of Computer Science & Engineering
University of Buffalo, SUNY
wjin2@buffalo.edu

Rohini K. Srihari
Dept. of Computer Science & Engineering
University at Buffalo, SUNY
rohini@cedar.buffalo.edu

Abhishek Singh
Dept. of Computer Science & Engineering
University at Buffalo, SUNY
singh5@cedar.buffalo.edu

ABSTRACT

Hypothesis generation is a crucial initial step for making scientific discoveries. This paper addresses the problem of automatically discovering interesting hypotheses from the web. Given a query containing one or two entities of interest, our algorithm automatically generates a semantic profile describing the specified entity or provides the potential connections between two entities of interest. We implemented a prototype on top of the Google search engine and the experimental results demonstrate the effectiveness of our algorithms.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: search process

General Terms: Algorithms, Experimentation

Keywords: web search, link analysis, hypothesis generation

1. INTRODUCTION

Finding the information about an entity or the relationships between two entities on the Web is a novel and challenging problem. Existing Web search engines excel in keyword matching and document ranking, but they cannot well handle such queries. In this paper, we treat the Web as a huge text database, and attempt to find the information concerning a specified entity and discover relationships between two entities of interest from unstructured documents. A traditional search involving, for example, two person names will attempt to find documents mentioning both of these individuals. This research focuses on a different interpretation of such a relationship query: what is the best connection and evidence across multiple documents that may potentially connect these two persons? For example two individuals may not co-occur in any single document, but they can still share some fact that they live in trailer parks (even though not the same trailer park) in different documents. This information is gleaned from multiple documents and existing Web search engines cannot help much in finding the answer. In this paper, we propose a method to handle such a query assuming that one or more instances of both entities occur in the corpus, but not necessarily in the same document.

2. RELATED WORK

Much of the work in hypotheses generation makes use of an idea originated by Swanson in the 1980s. He proposed a simple "A influences B, and B influences C, therefore A may influence C" model that is commonly referred as Swanson's ABC model [1]. Built within this discovery framework, [2, 3] used domain-specific knowledge and the MEDLINE database to find relationships between two biomedical entities. However, those methods proposed in [2, 3] do not work for general relationship queries in general text collections.

[4] presented a model of using Stepping Stones and Pathways (SSP) to connect topics in document collections. However, the proposed method has some limitations when working with Web pages. For example, that method first forms a query *Q* that contains both entities and uses *Q* to retrieve an intermediate document set from the collection. In the case of a Web search engine, this may lead to the situation that either no document is returned or all returned documents are related to a single entity. Moreover, that method cannot discover useful but non-obvious information across multiple documents.

Recently [5] proposed a new method for answering relationship queries on two entities on the Web. It generates an ordered list of Web page pairs. Each Web page pair consists of one Web page for either entity and the potential connecting terms capturing the relationships are identified. The limitation of this approach is that it can only find one level of intermediate connecting terms (i.e. $A \rightarrow B \rightarrow C$), whereas our method provides a mechanism that supports generating a multi-level path (i.e. $A \rightarrow B_1 \rightarrow \cdots \rightarrow B_n \rightarrow C$), which is extremely useful when there is very little information available between entities, in which case the associations between them are often more than one level of transitivity. Second, the method proposed in [6] needs the searcher to further analyze these potential connecting terms to find exact answers to queries. In contrast, this postprocessing step is coupled with our algorithm to facilitate interpretation of the importance of linking terms and exact answers to queries are provided to the searcher directly.

3. ANSWERING SINGLE ENTITY QUERY

We introduce *Entity Profile (EP)* to answer single entity queries and *EP* is built by first obtaining Web pages *W* using Entity *E* as the query term from a Web search engine (our experiments use Google), then reducing noise from obtained Web pages (This involves HTML comments, JavaScript code, tags removal and traditional IR based preprocessing steps such as stopwords removal), then identifying a relevant set of passages (our experiments use sentences) from the cleaned documents and then identifying characteristic terms from these passages and assessing their relative importance as descriptors of entity *E*. We assume the most useful information is typically centered around the entity and use windowing (e.g. sentences) to obtain this information (This is necessary because of the large amount of noise in the returned Web pages). Formally, the entity profile is presented as a weighting vector of all unique terms that occur in the local context of Entity *E*. We denote the set of passages in *W*

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

containing E by $S(E)$ and the size of the set by $|S(E)|$. *Profile (E)* is automatically learned and defined as follows:

$$\Pr ofile(E)=\{\omega_{e,1}c_1,\omega_{e,2}c_2,\cdots,\omega_{e,k}c_k,\cdots\} \tag{1}$$

Where c_k occurs in the local context of Entity E and $\omega_{e,k}$ represents the relative importance of term c_k as a descriptor of entity E and is defined as follows:

$$\omega_{e,k}=\log(1+P(c_k\mid E)\omega(c_k)) \tag{2}$$

$P(c_k|E)$ can be thought of as the probability that term c_k co-relates with E, i.e.,

$$P(c_k\mid E)=\frac{N(c_k,E)}{|S(E)|} \tag{3}$$

Where $N(c_k, E)$ represents the number of times term c_k and entity E co-occur in passages across W. $\omega(c_k)$ is the impact weight for term c_k when generating entity profiles. We assume that the terms also appearing frequently in $W\text{-}S(E)$ are less important, thus,

$$\omega(c_k)=\log(\frac{1+|N|-|S(E)|}{1+|N(c_k)|-N(c_k,E)}) \tag{4}$$

Where $|N|$ is the number of passages in W, $|N(c_k)|$ is the number of passages containing term c_k.

4. ANSWERING RELATIONSHIP QUERY

Given two entities of interest, the relationship query is handled by first retrieving from the Web search engine the top N relevant web pages for either entity. After cleaning and preprocessing steps, the union of documents retrieved for either entity respectively is represented as an undirected labeled graph where nodes represent unique terms occurring in the texts and edges denote the proximity relationships between terms in passages (e.g. sentences). Edge weights represent the strength of relationships between adjacent terms and are calculated based on the similarity between their respective contexts. Formally, for each node (including query nodes) in the graph, an entity profile is constructed using the method described in the previous section and weights are then calculated by $S_{i,j}$ defined below.

Definition: Let $\overrightarrow{ep(c_i)}$ and $\overrightarrow{ep(c_j)}$ be two entity profile vectors for term c_i and c_j respectively. Further, let $\vec{S}=(S_{i,j})$ be a scalar similarity matrix. Then, each $S_{i,j}$ is defined as:

$$S_{i,j}=\frac{\overrightarrow{ep(c_i)}\bullet\overrightarrow{ep(c_j)}}{\left|\overrightarrow{ep(c_i)}\right|\times\left|\overrightarrow{ep(c_j)}\right|} \tag{5}$$

The relationship query is then answered by retrieving and ranking the potential paths connecting entities of interest in the graph. The process is similar to the algorithm proposed in [6] except that a graphic visualization of the new transitive associations discovered together with the direct association is provided here which improves the understanding of the interaction between entities and their strongly correlated terms.

5. EXPERIMENTS

We have implemented a prototype of the proposed techniques on top of the Google search engine. Some preliminary results are shown below where the top 5 documents for one entity or either entity (in relationship queries) were retrieved from Google. Table 1 shows a portion of entity profile generated for '*Bin ladin'*. Figure 1 presents a partial graphic visualization of the relationships between entities '*Bush*' and '*Bin Ladin*' with a maximum allowable path length of 3. We also observe that some links (e.g. Al-qaeda – Afghanistan) are captured by different passages in different documents.

Table 1. Entity profile for 'Bin Ladin' (Top 6 Terms)

Saudi	0.809565
Afghanistan	0.720114
Sudan	0.663566
Islamic	0.679435
Taliban	0.646188
Al-qaeda	0.626979

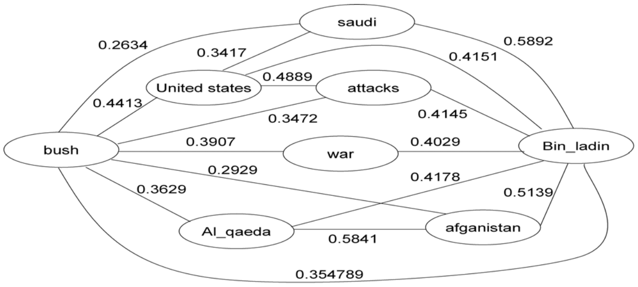

Fig. 1. Relationships between "Bush" and "Bin Ladin"

6. CONCLUSIONS

This paper addresses the problem of generating hypotheses from the web. We propose a method of generating entity profiles to answer single entity queries and exploring potential linkages via ranked intermediate terms when answering relationship queries. The method can also find cross-document links where existing Web search engines may not help much in finding the answer, and these non-obvious links across multiple documents may help to discover knowledge for generating hypotheses. These advantages are verified through a prototype implementation of our method on top of the Google search engine.

7. REFERENCES

[1] Swason, D. R. Complementary Structures in Disjoint Science Literatures. *ACM SIGIR (1991),* 280-289.

[2] Smalheiser, N. R. The Arrowsmith Project: 2005 Status Report. *Discovery Science* 2005: 26-43.

[3] Srinivasan, P. Text Mining: Generating Hypotheses from MEDLINE. *JASIST* 55(5): 396-413, 2004.

[4] Das-Neves, F., Fox, E. A. and Yu, X. Connecting Topics in Document Collections with Stepping Stones and Pathways. *CIKM '05*, ACM Press, Bremen, Ger., Nov. 2005, pp. 91-98.

[5] Luo, G., Tang, C. and Tian, Y. Answering Relationship Queries on the Web. *WWW2007*, pp. 561-570.

[6] Jin, W., Srihari, R. K., Ho, H. and Wu, X. "Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques". *(ICDM'07),* pp.193-202.