Improving Knowledge Discovery By Combining Text-Mining And Link-Analysis Techniques

Moty Ben-Dov
School of computing Science Middlesex University
The Burroughs, London NW4 4BT
UK

Wendy Wu
School of Computing Science Middlesex University
The Burroughs, London NW4 4BT
UK

Ronen Feldman
Department of Computer Science
Bar Ilan University
Ramat Gan
Israel

Paul A. Cairns
University College London
Remax House
31-32 Alfred Place
London, WC1E 7DP
UK

1. Abstract

The availability of online text documents exposes the readers to a vast amount of potentially valuable information buried in those texts. The huge number of documents created the pressing need for automated methods of discovering relevant information without having the need to read it all. Information Extraction (IE) from documents is one of approaches in text mining which extracts the features (entities) from documents.

We propose to use link-analysis techniques over the extracted features for finding new knowledge. In order to use link analysis techniques we need to create links out of the features extracted by the IE process.

In this paper we compare two approaches for creating links out of features; one is the co-occurrence approach and the other is the semantic approach. We tested both approaches and found that the semantic approach is preferable for finding focused information. For greater coverage of information we should use the Co-occurrence approach.

2. Introduction

"*Knowledge can be created by drawing inference from what is already known*" (Davies 1989). In his article, Davies explained how we could create new knowledge simply by knowing how to search the known information. The principle of his concept is based on the hypothesis that "*the wealth of recorded knowledge is greater than the sum of its parts"(Davies 1989).* This concept was the base for some research in different scientific fields. Swanson and Smalheiser (Swanson and Smalheiser 1999) found in their project that new interesting and unknown implicit information could be discovered if we study the linkage between textual records. Small and Garfield (Small and Garfield 1989) suggested a wider citation search. They suggested that a path of several co-citation steps can connect two clusters of documents. Therefore, it will be useful to study the multistage paths between topics when we are forming hypotheses.
The CMU group (Craven, DiPasquo et al. 2000) found that new knowledge could be learned from hyperlink paths on the WWW.

The examples above reflected the fact, that we could discover new knowledge from existing information base, if we can assemble the pieces of existing knowledge in the right way. New concepts could be formed, if we could use some techniques to discover previously unknown logical connections among the existing information we have.

The main theme of this paper is that **new knowledge may be discovered in a document collection by combining text-mining and link-analysis techniques.**

Link-analysis is the process of building up networks of interconnected objects in order to explore pattern and trends. Link-analysis is based on a branch of mathematics called "graph theory"(Barry and Linoff 1997; Wastphal and Blaxton 1998).

The known link-analysis tools (such as NETMAP, Visual Analytics) operate only on structure data. The links between various entities need to be feed exclusively. They can not work on unstructured data such as text. To overcome this obstacle we used an IE tool which extracted the features from

the text. Those features were then used as input in our link analysis & text mining system.

In this paper we will compare two techniques for creating links out of features (entities that were extracted by IE process), namely co-occurrence link and semantic link, which will be explained in details in the next section.

We defined three questions and checked the quality of the answers we get from each technique. The answer will be based on the quality of the links each technique created.

We will make the comparison by calculating the precision and recall of the links each technique created.
The comparison will be used to emphasis the strength and weakness of each technique compared to each other.

3. Foundations

3.1. Co-occurrence links

The co-occurrence links were created by a simple method of seeking the existence of the relevant features within the same sentence. This method was implemented by using a pattern matching mechanism without referring to any syntactic or semantic role (Feldman, Regev et al. 2002).

Two features co-occur within a sentence if they both are contained in it. Figure 1 is an example of Relation Map of the co-occurrence relations between persons in our data. The darker the relation between two persons, the more links their co-occurrence relations contain. In this graph the threshold for showing co-occurrence relation between two persons was set to 10 links.

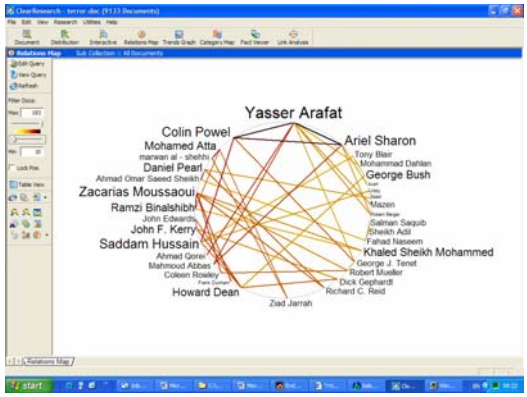

Figure 1. Co-occurrence relations between persons.

3.2. Semantics links

The semantic links were created by using noun phrase and verb identification and linguistic and semantic constraints. In this process the tool extracts predefined semantic relationships by performing a deep syntactic and semantic analysis of the sentence. The process uses a strategy of 5 layers (Feldman, Regev et al. 2002):

- Layer 0 – part of speech tagger
- Layer 1 – noun phrase and verb phrase grouper
- Layer 2 - Verb and Noun Pattern Extractor
- Layer 3 - Named Entity Recognizer
- Layer 4 - Template ('relationship') extractor

The implementation of the semantic analysis was done by using the rule base general IE language DIAL (Declarative Information Analysis Language). This language was developed at ClearForest® labs (Feldman, Regev et al. 2002).

Figure 2 is an example of the semantic relations between persons in our data. In this graph, the threshold for showing semantic relation between two persons was set to 3 links.

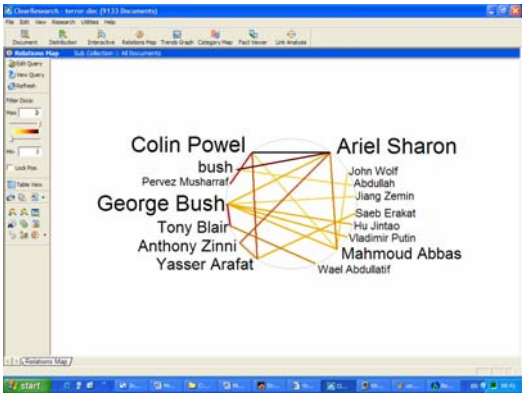

Figure 2. Semantic relations between persons.

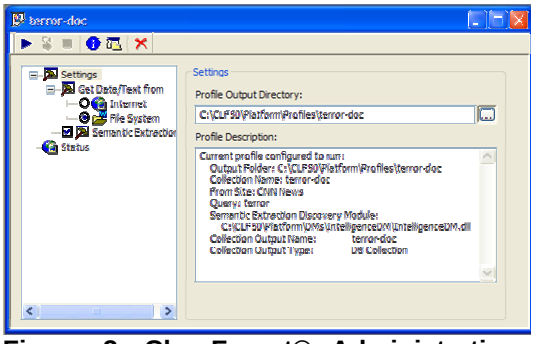

Figure 3. ClearForest® Administration tool for collecting and processing the data

4. Comparison Experiments

4.1. The tools

We used the ClearForest™ suite as a platform for the research. From the suite we used following tools:

- Administration tool™ (5.0) – we use it for data collection and semantic features extraction.
- ClearResearch™ (5.0) – we used it for data analysis and visualization.

4.2. The Data

We created a database of web news pages from four news sites – CNN, BBC, CBS and Yahoo. The search criterion for the SEDP (Semantic Extraction Discovery Process), which is depicted in Figure 3, was "terror". Each page was downloaded from the web, modified and passed a SEDP. ClearForest® Administration tool extracts features from the pages by using a predefined IE Rule Base model for Intelligence (Feldman, Liberzon et al. 2000). At the end of this process a data-base of features extracted from 9133 web pages and taxonomy that were generated. During the SEDP, the co-occurrence and semantic links were also created.

4.3. The experiments

In our experiments we will compare two approaches for creating links, the co-occurrence and the semantic. We will find the links each approach had created and then calculate and compare the precision and recall of each approach.

We searched information about meetings between persons by using the links we created. We checked if we can find answers for the following questions:

Q1. How many meetings did Arial Sharon and Colin Powel have?
Q2. How many meetings did Yasser Arafat and Colin Powel have?
Q3. How many meetings did Yasser Arafat and Anthony Zinni have?

The above questions were answered by using co-occurrence links and semantics links respectively.

We will explain the process of getting answers to Q1 as an example.

4.3.1. Preliminary stage

The preliminary stage, of the process, was to find the number of all the documents that mention meetings between Ariel Sharon and Colin Powel in the documents we collected. We did a query for all the documents in which Sharon and Powel appear. We found 183 such documents. We read all of the 183 and 22 of them were about meetings between Sharon and Powel.
The result of this stage was that the total number of correct links in the database is 22.

4.3.2. Co-occurrence links

In the first stage of the experiment, we choose the co-occurrence links, at the sentence level, between Ariel Sharon and Colin Powel (Figure 4).

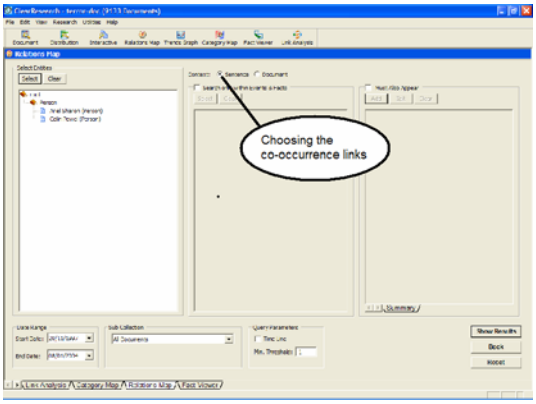

Figure 4. Choosing the co-occurrence links with ClearResearch™

We got 71 documents that mentioning a meeting between Sharon and Powel which contain at least one sentence with both names (Figure 5). We found that from these 71 documents there were 20 sentences about meetings between Sharon and Powel (each such sentence created a link).

The results of this stage were; the number of correct links is 20 and the number of total links is 71.

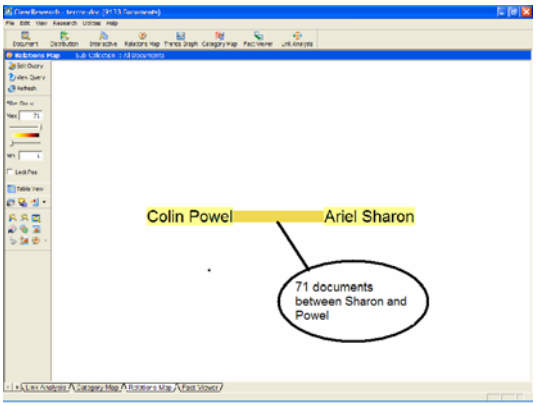

Figure 5. The Co-occurrence links between Colin Powel and Ariel Sharon

4.3.3. Semantic links

The next stage in the research was to check the semantic links. We checked the **Person_Meeting** events that the SEDP had found (Figure 6). The semantic meaning of this event is that the text fragment includes a reference to an actual meeting between two persons. The following example explains the SEDP event extraction: In document no 5578 we found the sentence:

"Powell met earlier with Israeli Prime Minister Ariel Sharon to discuss how Israel might end its military operation in Palestinian cities"

The SEDP process will extract the following XML sentence which indicated **Person_Meeting** event (link):

<**Person_Meeting**>
<**Person**>*Powell*<**Person**>
<**Meeting**> *met* <**Meeting**>
<**Person**> *Ariel Sharon* <**Person**>
<**Person_Meeting**>

We found that 9 documents have the semantically **Person_Meeting** links. After reading the 9 documents, 8 of them were about meetings between Sharon and Powel.

The results of this stage were; the number of correct links is 8 and the number of total links is 9.

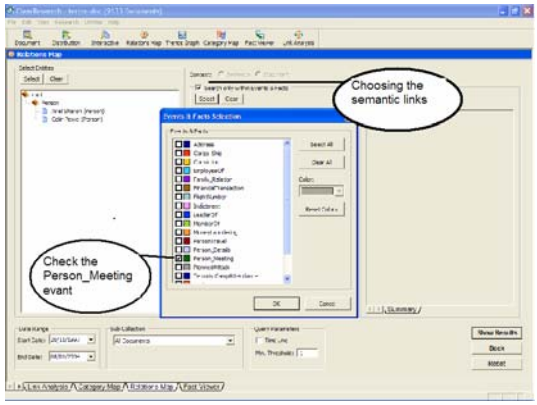

Figure 6. Choosing the semantic link Person_Meeting with ClearForest®

We pass the 3 stages experiment for the remaining questions (Q2, Q3) and the detailed results will be on the next section.

5. The Results

We calculated the Precision and the Recall of each phase in our research.

The **Precision** was calculated as the correct links (i.e. links that report an actual meeting) divided by the Total number of links found.

The **Recall** was calculated as the correct links that were found divided by the total

number of correct links on the on the whole database.

Table 1. The Q1 results

	Co-occurrence links	Semantic links
Correct links	20	8
Total Correct links in the database	22	22
Total links	71	9
Precision	28.17%	88.89%
Recall	90.91%	36.36%

Table 2. The Q2 results

	Co-occurrence links	Semantic links
Correct links	14	5
Total Correct links in the database	15	15
Total links	94	6
Precision	14.89%	83.33%
Recall	93.33%	33.33%

Table 3. The Q3 results

	Co-occurrence links	Semantic links
Correct links	8	5
Total Correct links in the database	11	11
Total links	9	5
Precision	88.89%	100%
Recall	72.73%	45.45%

6. Discussion and Conclusions

The links analysis process has three steps (Figure 7).

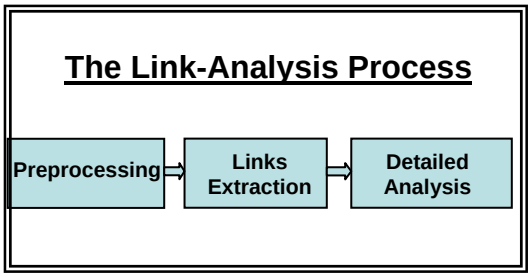

Figure 7. Link Analysis Process

The first step is the data preprocess. We need to preprocess the original text and the features that were extracted by the IE tool.
The co-occurrence links preprocessing is almost obvious. We need to do very little preprocessing if any.
On the other hand the preprocessing for semantic links is much more time consuming. We need to develop the extraction rules for each link type we want to use. Based on our experience it could take 1-3 weeks for an experience programmer to build a set of rules for each link type.

The second step is the actual links extraction, it's an automatic process. There is no difference, in cost, between the co-occurrence links and the semantic links.

The third step is the detailed analysis of the founded links.
With the semantic links the results are more focused and we can reach the needed information easily and quickly.
With the co-occurrence links we have to devote much more effort to reach the needed information since the noise level is quit high.

We found that the semantic approach is a time consuming process during the preprocessing step but we gain it at the analysis step. With the co-occurrence approach we found that the preprocessing step is a short step but the analysis step is time consuming step.

In the three experiments we did, the precision in the semantic links was significantly better than the precision we got with co-occurrence links. The results

were expected because the co-occurrence is a naïve technique for extracting accurate links.

The recall parameter shows the strength of the co-occurrence technique. With the co-occurrence technique, we got good coverage of the links we were looking for. We didn't have 100% recalls because the co-occurrence doesn't refer to anaphors of persons names. The following sentence wasn't recognized by using co-occurrence at the sentence level as a meeting between Powel and Sharon (document no 1653).
"***He*** *will be holding his first talks with Israeli Prime Minister* ***Ariel Sharon*** *and new Palestinian Prime Minister Mahmoud Abbas, known informally as Abu Mazen, since the road map was published.*"
The anaphora "He" refers to Powel but the co-occurrence process couldn't find it. This document was found by the preliminary step at the document level because the name Powel exists in other sentences in the document.

The poor results on the recall of the semantic links could be as a result of the SEDP we used. It was a general SEDP that was built for general intelligent semantic extraction. After learning the documents the SEDP missed we think we could add and modify some of the rules in the SEDP and considerably improve the recall of semantic links.

Our main conclusion is that if we need very focused information then the best results will be obtained by using the semantic links. When we look for greater coverage of information we should used the co-occurrence links.

In the future we plan to test a third technique for creating links, the statistical technique. We will extract links by using statistical information extraction techniques such as HMM (Seymore, McCallum et al. 1999).
Finally, we believed that the best results for links extraction will be achieved by a hybrid method of all the above techniques.

7. Actual usage of link-analysis in discovering new anti-terror knowledge

In this section we will demonstrate how we can use the co-occurrence links we created to discover new knowledge in the anti-terror domain.

Our initial query was "are there any relationships between Osama Bin Laden and John Paul II (The Pope)?"

We did a link-analysis research with the ClearResearch tool using the co-occurrence links we discovered with the SEDP. We found that there is no direct connection but we found an indirect connection between Bin Laden and the Pope thru Ramzi Yousef (Figure 8).

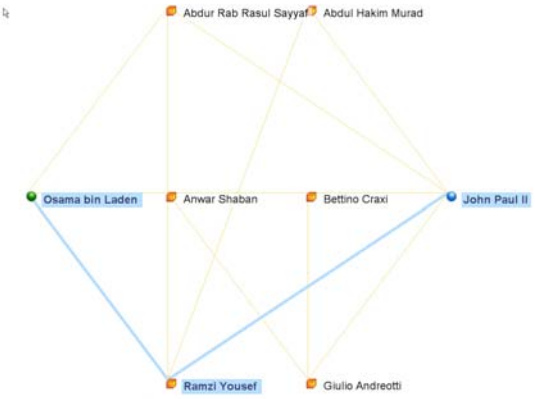

Figure 8 - Link Analysis, finding relationships between Osama Bin Laden and John Paul II

We looked at the documents which support the co-occurrence link. We found 6 such documents which connected Osama Bin Laden and Ramzi Yousef and highlighted the actual sentence within the documents (Figure 9).

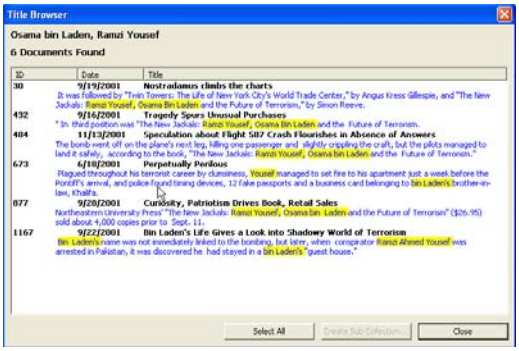

Figure 9 - documents supporting the connection between Osama bin Laden and Ramzi Yousef

We then looked at documents which supported the co-occurrence link between John Paul II and Ramzi Yousef and we found one such document (Figure 10).

Figure 10 - documents supporting the relationship between Ramzi Yousef and the Pope

After reading the documents we can understand that in 1995, Ramzi Yousef was responsible for the attempt to assassinate John Paul II when he was visiting Manila/Philippine.

We can also infer that Ramzi Yousef was part of Osama Bin Laden terror organization. So we can assume, with very high confidence, that Osama Bin Laden's organization was responsible for the assassination attempt of John Paul II.

The above is new knowledge which we discovered within our document collection by using link-analysis techniques.

8. REFERENCES

Barry, M. and G. Linoff (1997). "Data Mining Techniques - for marketing, sales and customer support." Wiley Computer Publishing: 216-242.

Craven, M., D. DiPasquo, et al. (2000). "Learning to construct knowledge bases from the World Wide Web." Artificial Intelligence **118**(1-2): 69-113.

Davies, R. (1989). "The creation of new knowledge by information retrieval and classification." Journal of Documentation **45**(4): 273-301.

Feldman, R., Y. Liberzon, et al. (2000). A framework for specifying explicit bias for revision of approximate information extraction rules. KDD**:** 189-197.

Feldman, R., Y. Regev, et al. (2002). "Mining biomedical literature using information extraction." Current Drug Discovery: 19-23.

Seymore, K., A. McCallum, et al. (1999). Learning Hidden Markov Model Structure for Information Extraction. AAAI 99 Workshop on Machine Learning for Information Extraction.

Small, H. and E. Garfield (1989). "Verification of results that logically related noninteractive literatures are potential source of new knowledge." Journal of the American Society for Information Science **40**(3): 152.

Swanson, D. R. and N. R. Smalheiser (1999). "Implicit text linkages between Medline records; using Arrowsmith as an aid to scientific discovery." Library Trends **48**(1): 48-59.

Wastphal , C. and T. Blaxton (1998). "Data Mining Solutions - Methods and tools for solving real-world problems." Wiley Computer Publishing: 201-264.