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

Presentation By Moty Ben-Dov

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

Dr. Wendy Wu
Middlesex University London

Dr. Ronen Feldman
Bar Ilan University Israel

Dr. Paul A. Cairns
University College London

The presentation framework

The reasons for combining text-mining and link-analysis

Two links extraction approaches

The experiments and the results

Discussion and Conclusions

The presentation framework

The reasons for combining text-mining and link-analysis

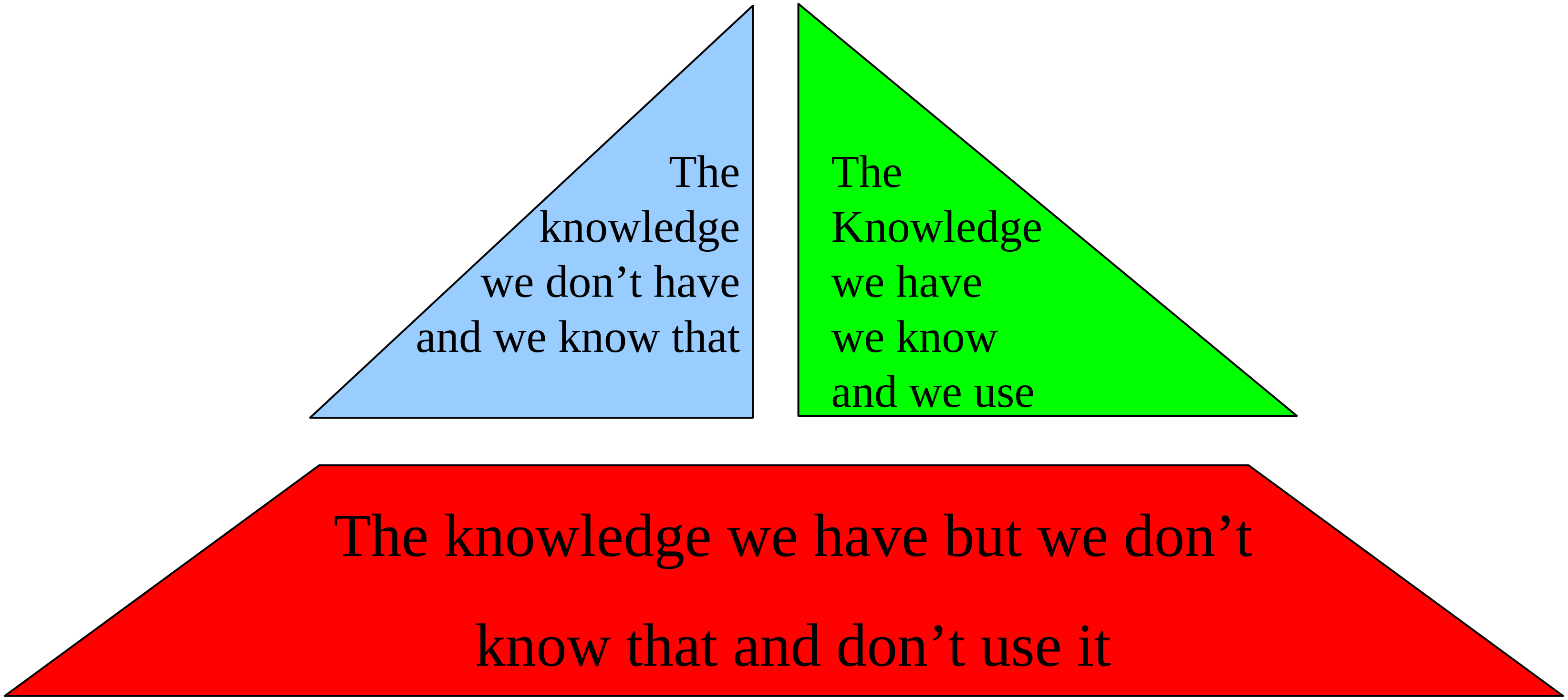
Two links extraction approaches

The experiments and the results

Discussion and Conclusions

The reasons for combining text-mining and link-analysis

The organization’s knowledge



The knowledge we don’t have and we know that

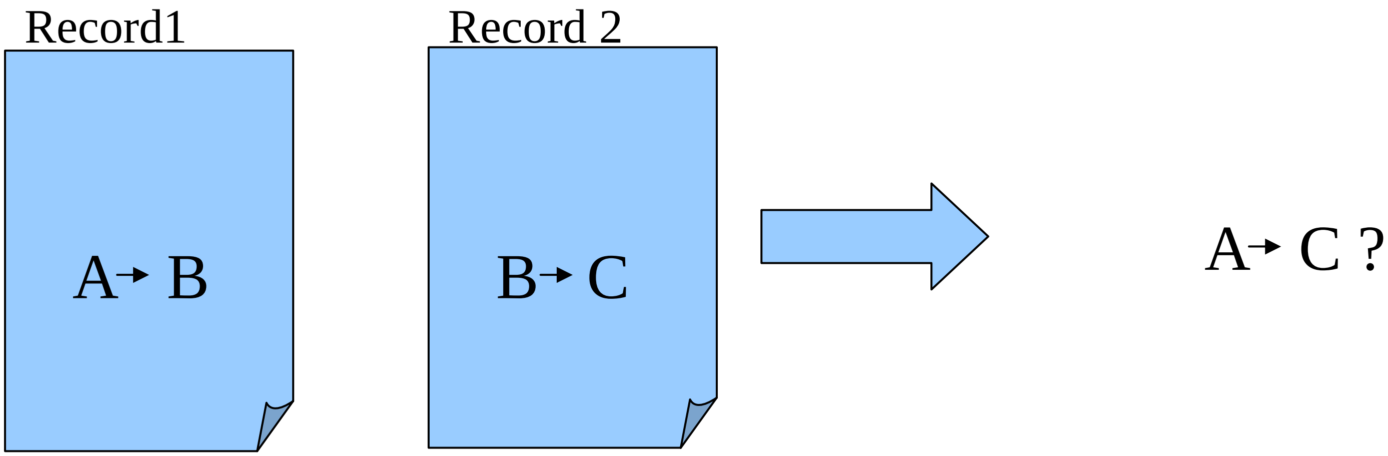
The Knowledge we have we know and we use

The knowledge we have but we don’t know that and don’t use it

The reasons for combining text-mining and link-analysis

”Knowledge can be created by drawing inference from what is already known”

Ron Davies 1989



Record1

A → B

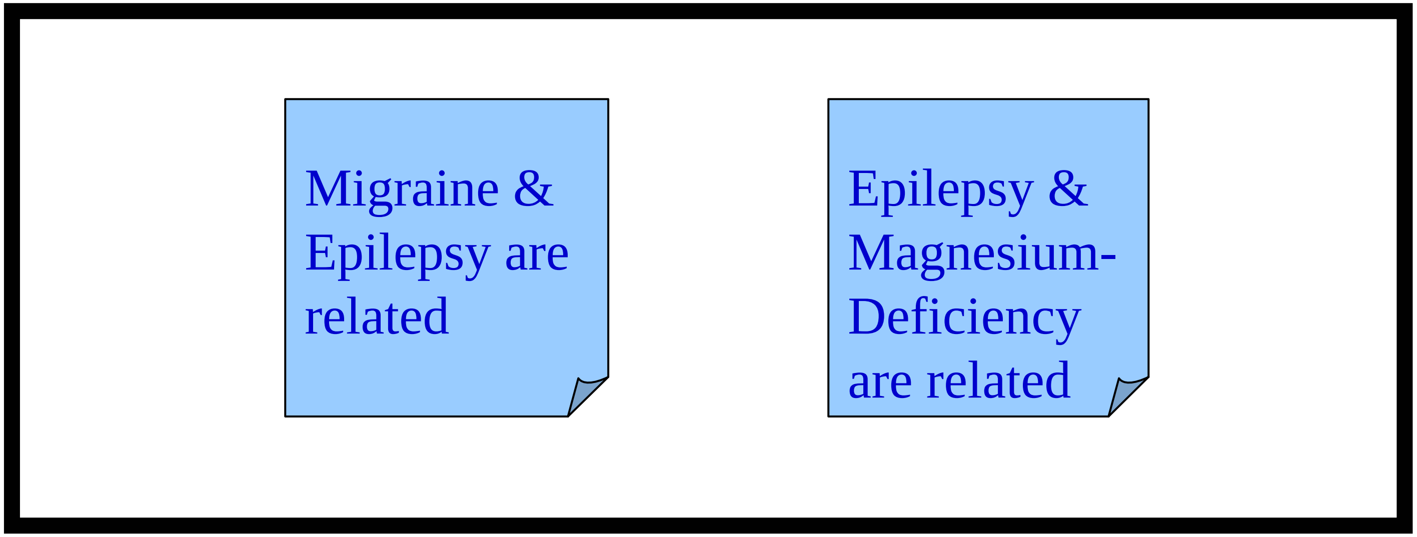
Record 2

B → C

A → C ?

The reasons for combining text-mining and link-analysis

Arrowsmith project (Swanson & Smalheiser)



Migraine & Epilepsy are related

Epilepsy & Magnesium-Deficiency are related

Migraine → Magnesium-Deficiency

The reasons for combining text-mining and link-analysis

What is Text-Mining (TDM)?

“TDM is the process of extracting interesting patterns from very large unstructured content database for the purposes of discovering knowledge.

TDM applies the same analytical functions used to do Data-Mining and also applies natural language (NL) and information retrieval (IR) techniques.”

Dörre, Gerstl and Seiffert 1999

The reasons for combining text-mining and link-analysis

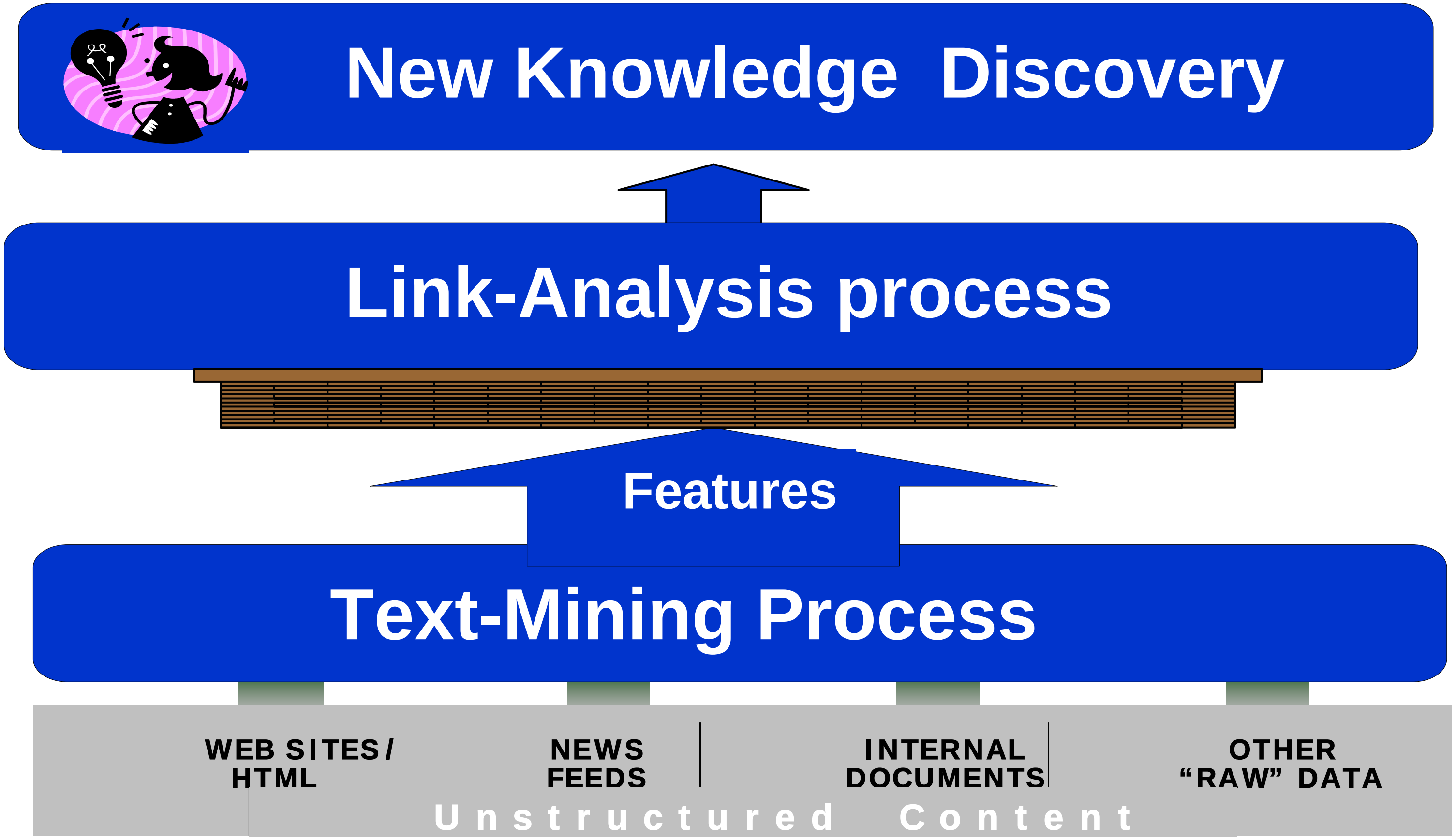
What is Link Analysis?

“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

The reasons for combining text-mining and link-analysis



New Knowledge Discovery

Link-Analysis process

Features

Text-Mining Process

WEB SITES/ HTML

NEWS FEEDS

INTERNAL DOCUMENTS

OTHER "RAW" DATA

Unstructured Content

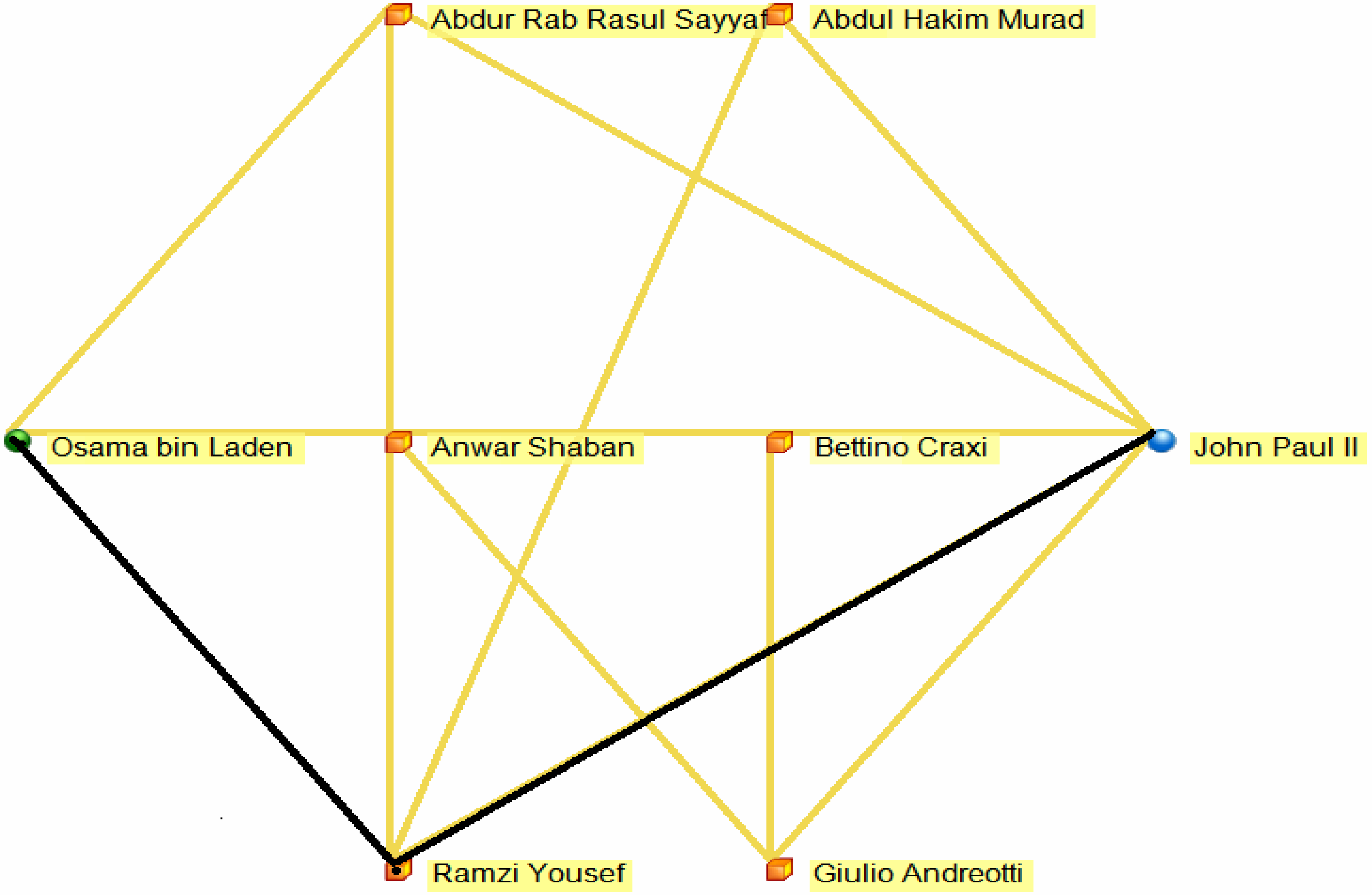
Improving Knowledge Discovery Process

The reasons for combining text-mining and link-analysis

Actual usage of combining text mining and link-analysis for discovering new anti-terror knowledge

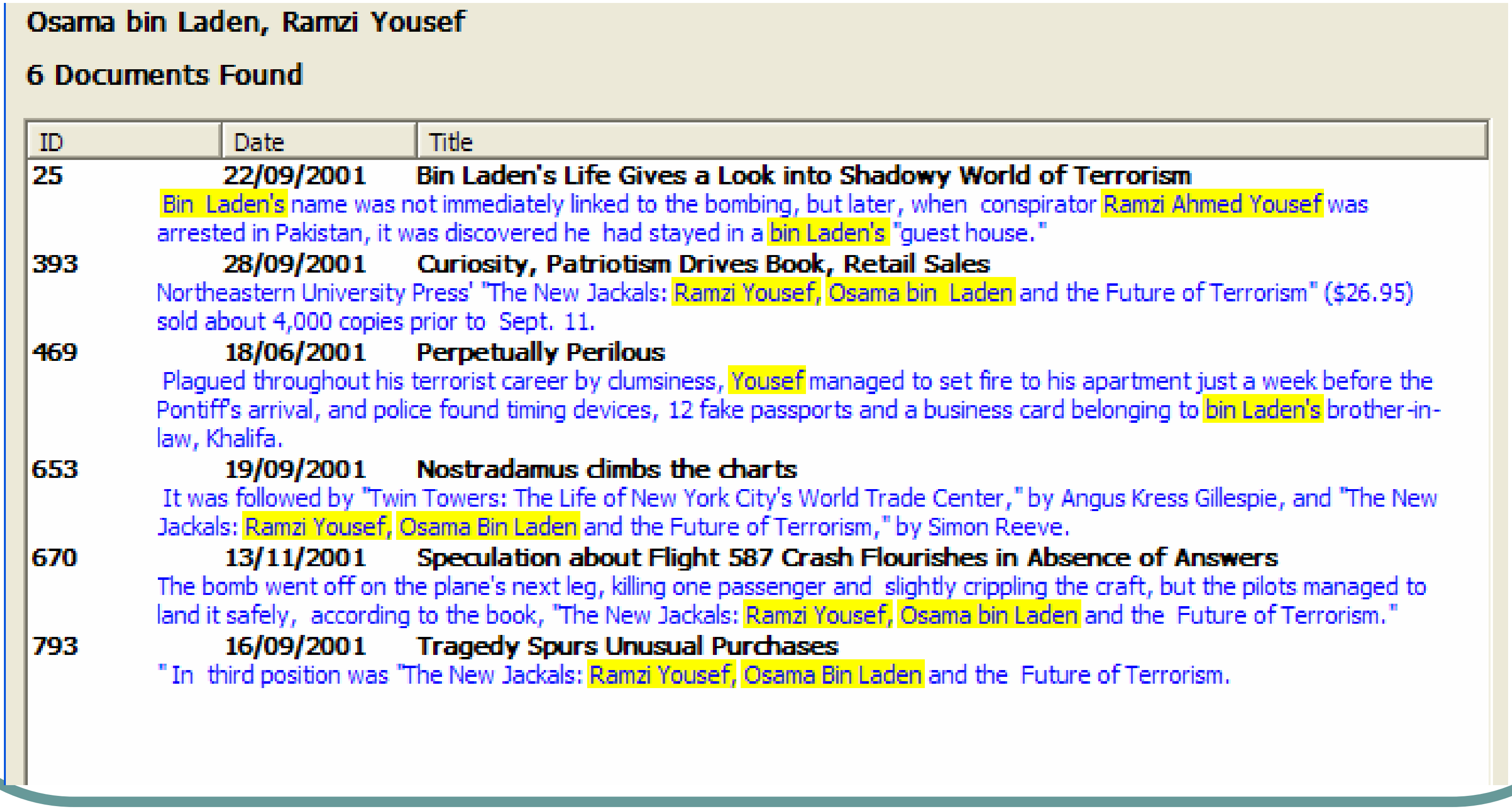
- The mission was to find a connection between “Bin Laden” and “John Paul II”.
- We ran a Text-Mining tool over a document database and extracted all persons names in the documents (features).
- We tried to find connection between the two by using just the features. No connection was found.
- The next step was to used a Link-Analysis tools to find indirect connections.
- For the Link-Analysis process we used the Co-occurrences links at the sentence level.

The reasons for combining text-mining and link-analysis



The reasons for combining text-mining and link-analysis

Documents supporting the connection between Osama bin Laden and Ramzi Yousef

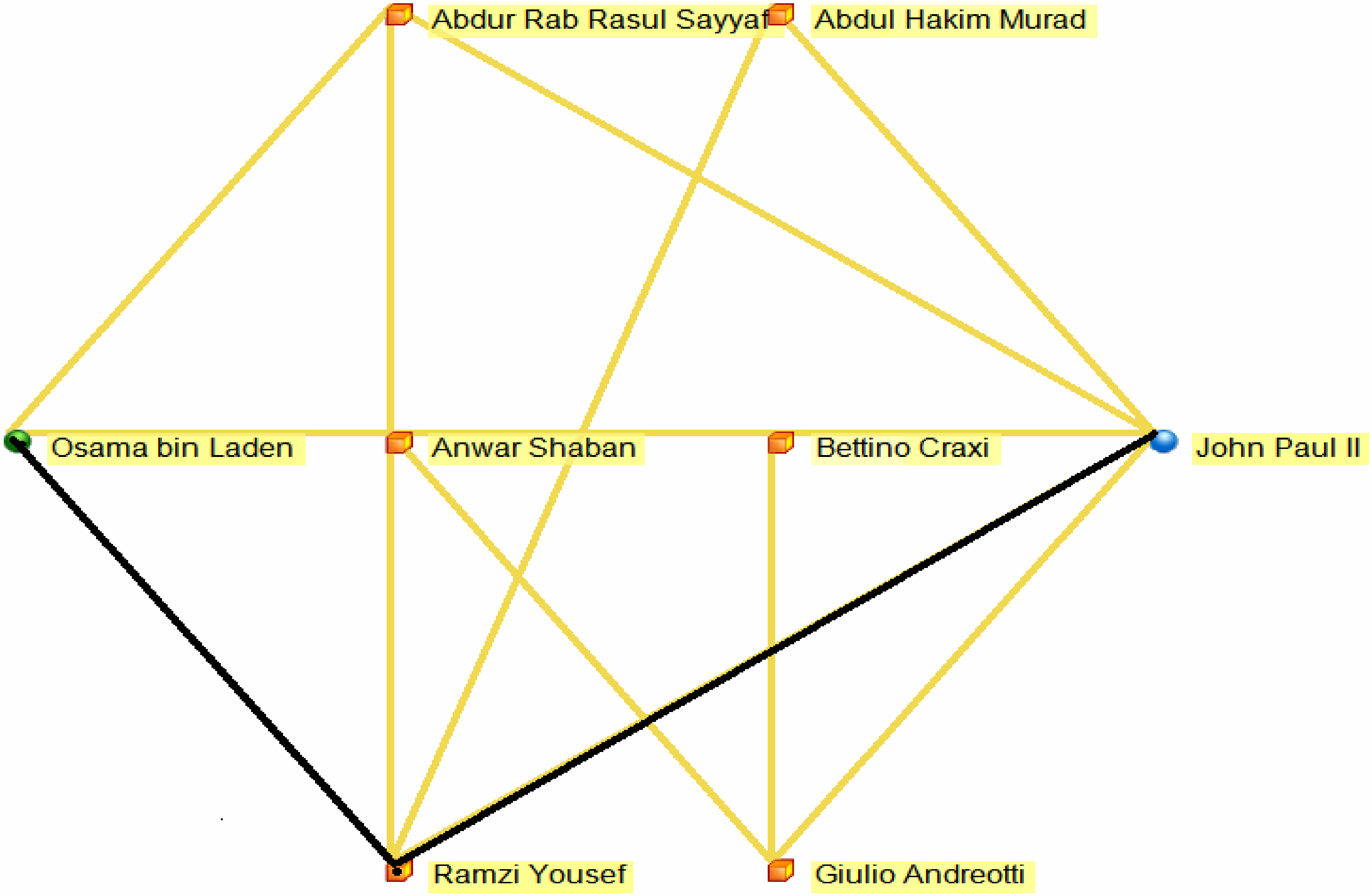


Osama bin Laden, Ramzi Yousef

6 Documents Found

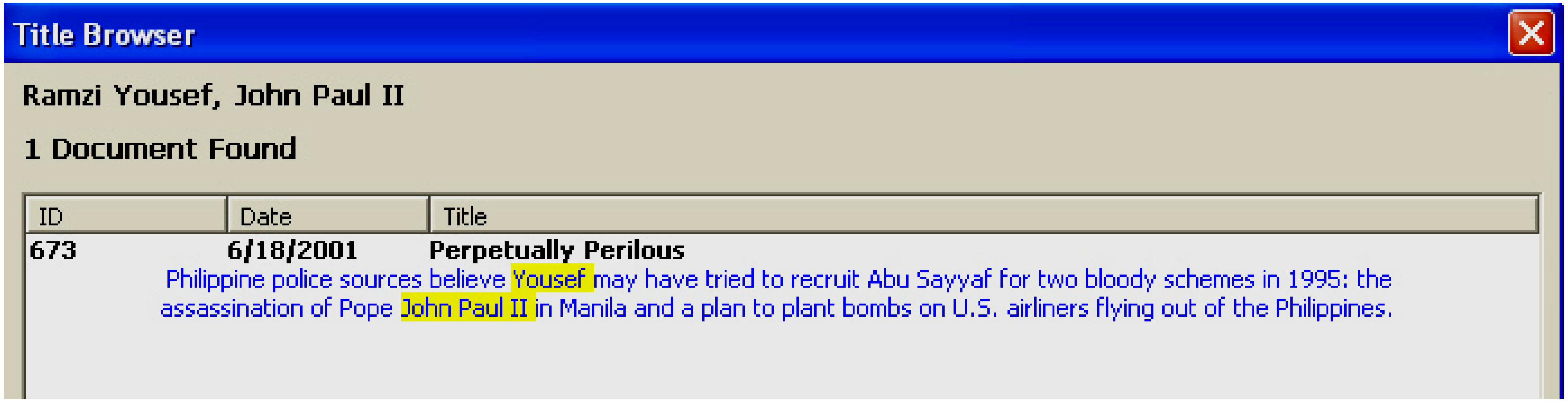
ID	Date	Title
25	22/09/2001	**Bin Laden's Life Gives a Look into Shadowy World of Terrorism** Bin Laden's name was not immediately linked to the bombing, but later, when conspirator Ramzi Ahmed Yousef was arrested in Pakistan, it was discovered he had stayed in a bin Laden's "guest house."
393	28/09/2001	**Curiosity, Patriotism Drives Book, Retail Sales** Northeastern University Press' "The New Jackals: Ramzi Yousef, Osama bin Laden and the Future of Terrorism" ($26.95) sold about 4,000 copies prior to Sept. 11.
469	18/06/2001	**Perpetually Perilous** Plagued throughout his terrorist career by clumsiness, Yousef managed to set fire to his apartment just a week before the Pontiff's arrival, and police found timing devices, 12 fake passports and a business card belonging to bin Laden's brother-in-law, Khalifa.
653	19/09/2001	**Nostradamus climbs the charts** It was followed by "Twin Towers: The Life of New York City's World Trade Center," by Angus Kress Gillespie, and "The New Jackals: Ramzi Yousef, Osama Bin Laden and the Future of Terrorism," by Simon Reeve.
670	13/11/2001	**Speculation about Flight 587 Crash Flourishes in Absence of Answers** The bomb went off on the plane's next leg, killing one passenger and slightly crippling the craft, but the pilots managed to land it safely, according to the book, "The New Jackals: Ramzi Yousef, Osama bin Laden and the Future of Terrorism."
793	16/09/2001	**Tragedy Spurs Unusual Purchases** "In third position was "The New Jackals: Ramzi Yousef, Osama Bin Laden and the Future of Terrorism.

The reasons for combining text-mining and link-analysis



The reasons for combining text-mining and link-analysis

Documents supporting the relationship between Ramzi Yousef and the Pope



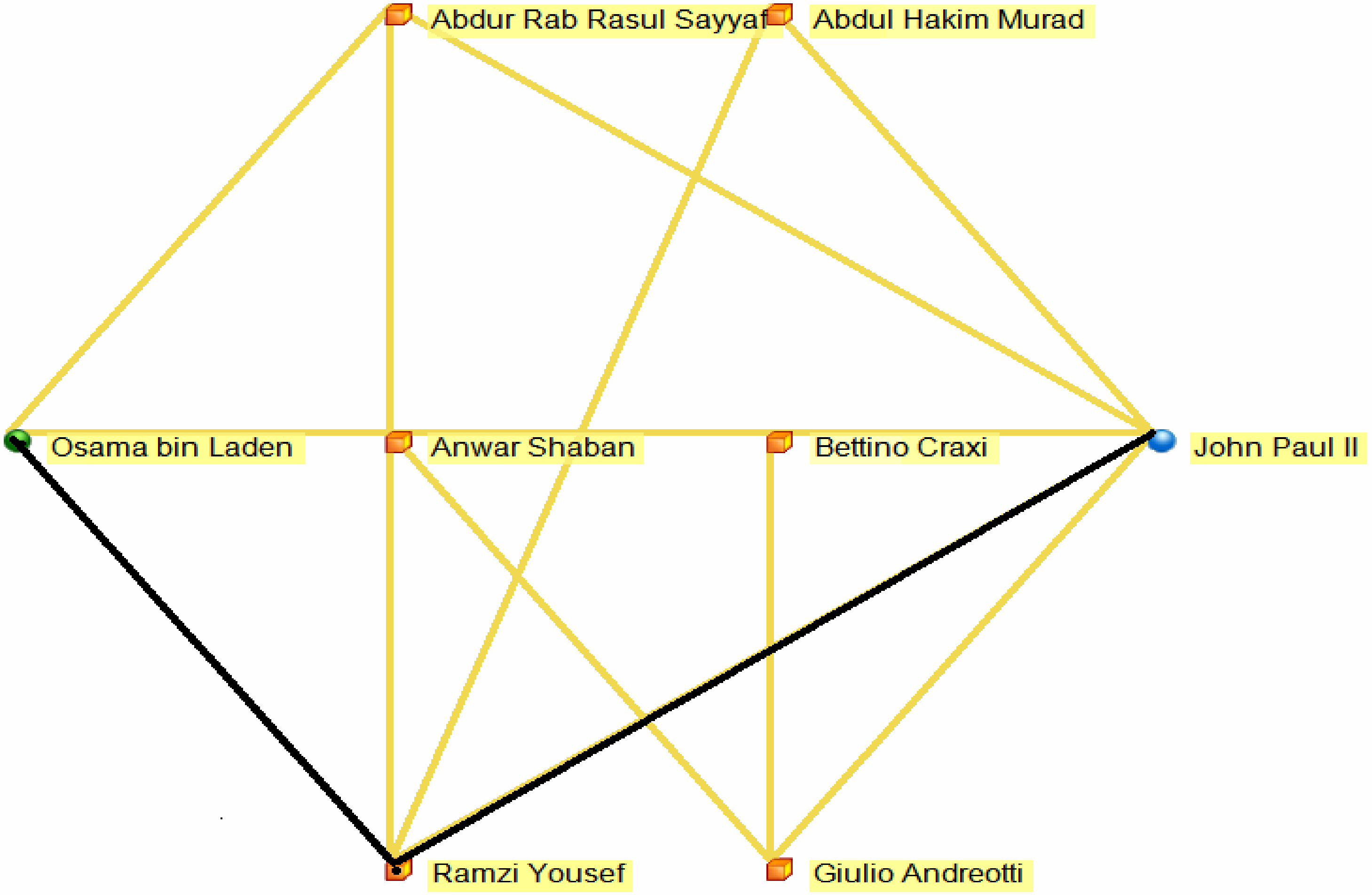
Ramzi Yousef, John Paul II

1 Document Found

ID	Date	Title
673	6/18/2001	**Perpetually Perilous**

Philippine police sources believe Yousef may have tried to recruit Abu Sayyaf for two bloody schemes in 1995: the assassination of Pope John Paul II in Manila and a plan to plant bombs on U.S. airliners flying out of the Philippines.

The reasons for combining text-mining and link-analysis



The presentation framework

- The reasons for combining text-mining and link-analysis
- Two links extraction approaches
- The experiments and the results
- Discussion and Conclusions

Two links extraction approaches

Co-occurrence links **-** Two features co-occur within a sentence if they both appear in the same sentence

the co-occurrence links were created by a simple method of seeking the existence of the relevant features within the same sentence

Semantics links - The semantic links were created by using noun phrase ,verb identification ,linguistic and semantic constraints.

The presentation framework

The reasons for combining text-mining and link-analysis

Two links extraction approaches

The experiments and the results

Discussion and Conclusions

The experiments and the results

The experiments Framework

The target **-** to compare between the two links extraction methods Co-occurrence and Semantic.

The Data – 9133 web pages about “*terror*” from the CNN, CBS, BBC and Yahoo.

The Tools – we used the ClearForest® suite for data collection, Text mining (features extraction), link extraction and results visualization.

The experiments and the results

The experiment Framework (cont.)

- We searched information about meetings between persons by using the links each approach had 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?

- For each question we counted the meetings each approach had found and then calculate and compare the precision and recall of each approach.

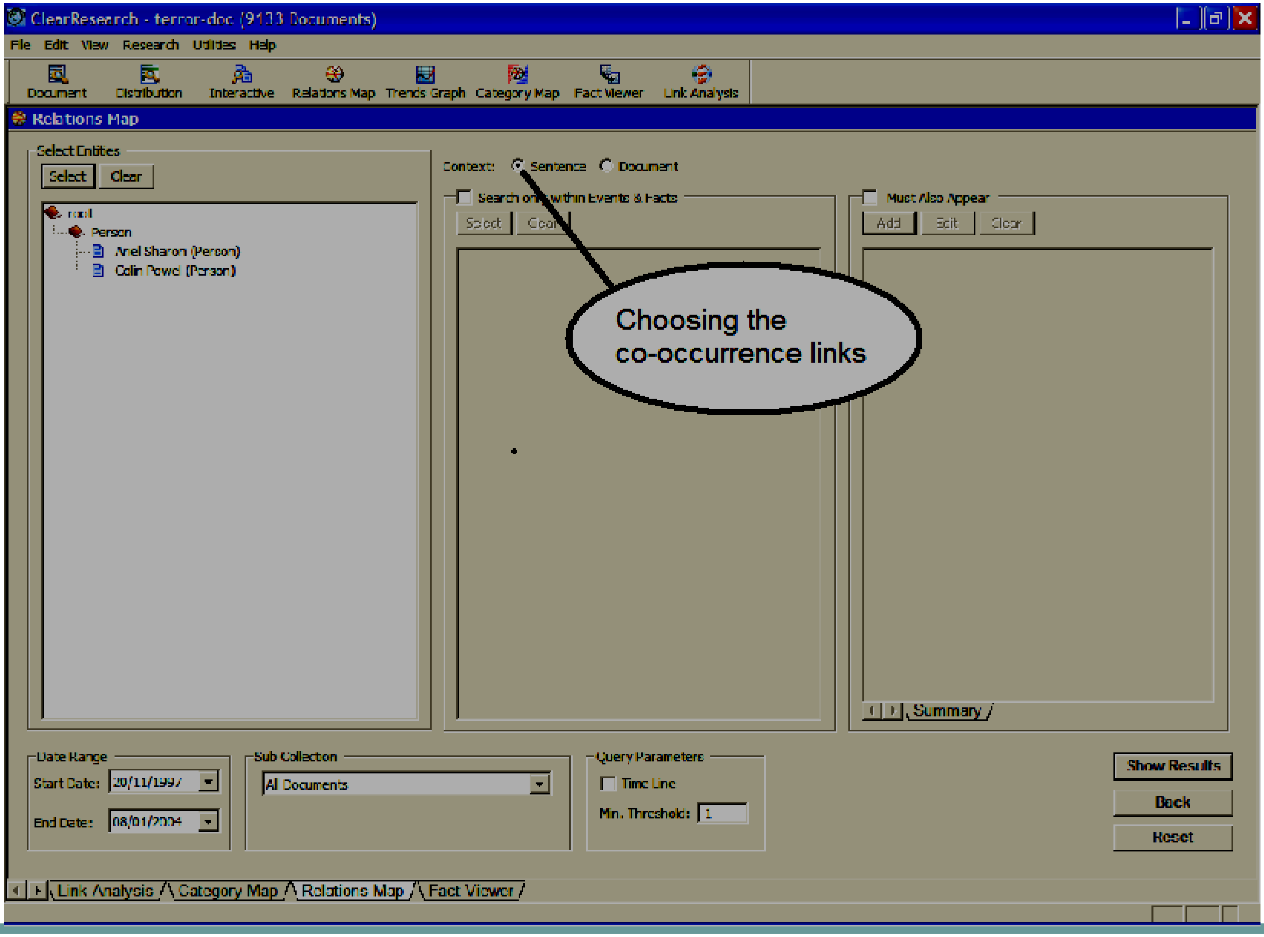
The experiments and the results

The experiment

For the first question Q1 (How many meetings did Arial Sharon and Colin Powel have?) we did the following steps :.

1. **Preliminary stage** – finding the actual number of documents that mention meetings between *Ariel Sharon* and *Colin Powel*. We did a query for all the documents in which Sharon and Powel appear and found 183. we read the documents and found 22 documents that mention a meeting between them.
2. **The co-occurrence links** –We chose the co-occurrence at the sentence level and we got 71 sentences (links). We found that only 20 links are actually about a meeting.

The experiment (cont.)



root
- Person
  - Ariel Sharon (Person)
  - Colin Powell (Person)

Context: ● Sentence ○ Document

□ Search only within Events & Facts

□ Must Also Appear

Date Range

Start Date: 20/11/1997

End Date: 08/01/2004

Sub Collection: All Documents

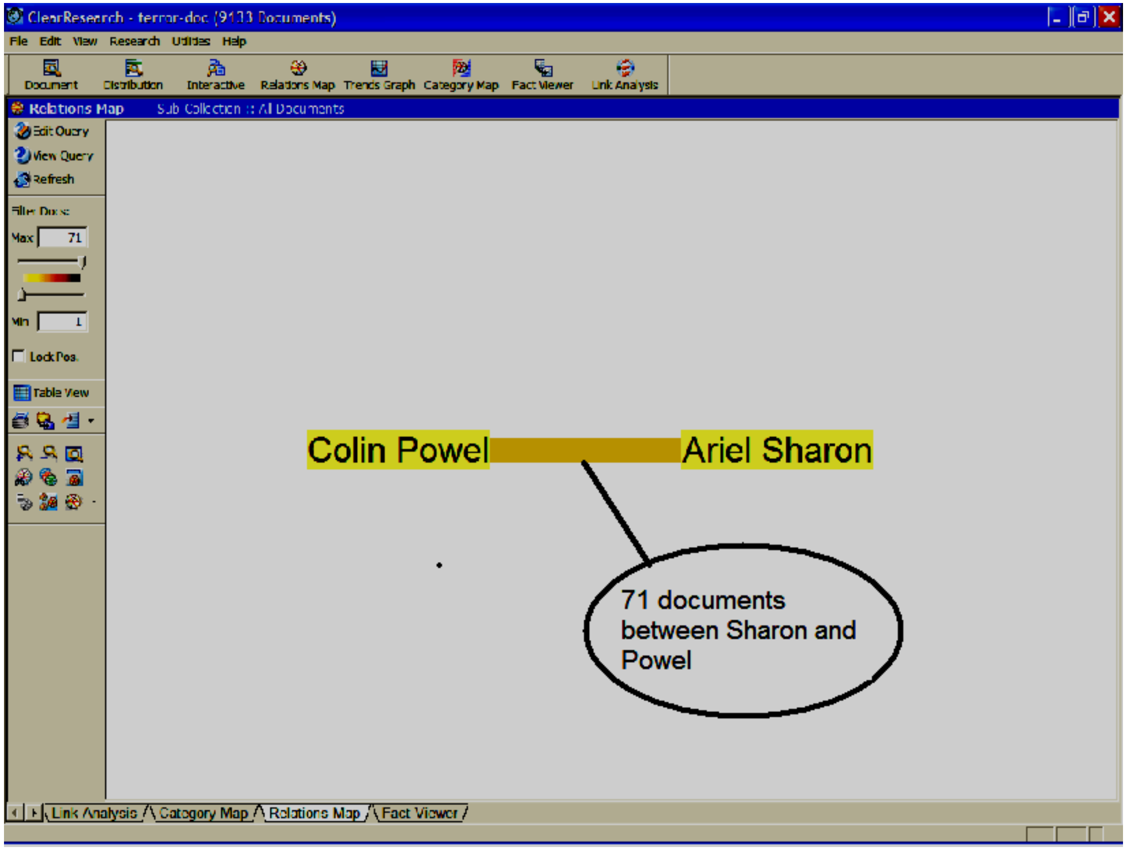
Query Parameters

□ Time Line

Min. Threshold: 1

Choosing the co-occurrence links

The experiment (cont.)



71 documents between Sharon and Powel

The experiments and the results

The experiment (cont.)

3. **The Semantic links** – The links were created by using noun phrase and verb identification and linguistic and semantic constraints.

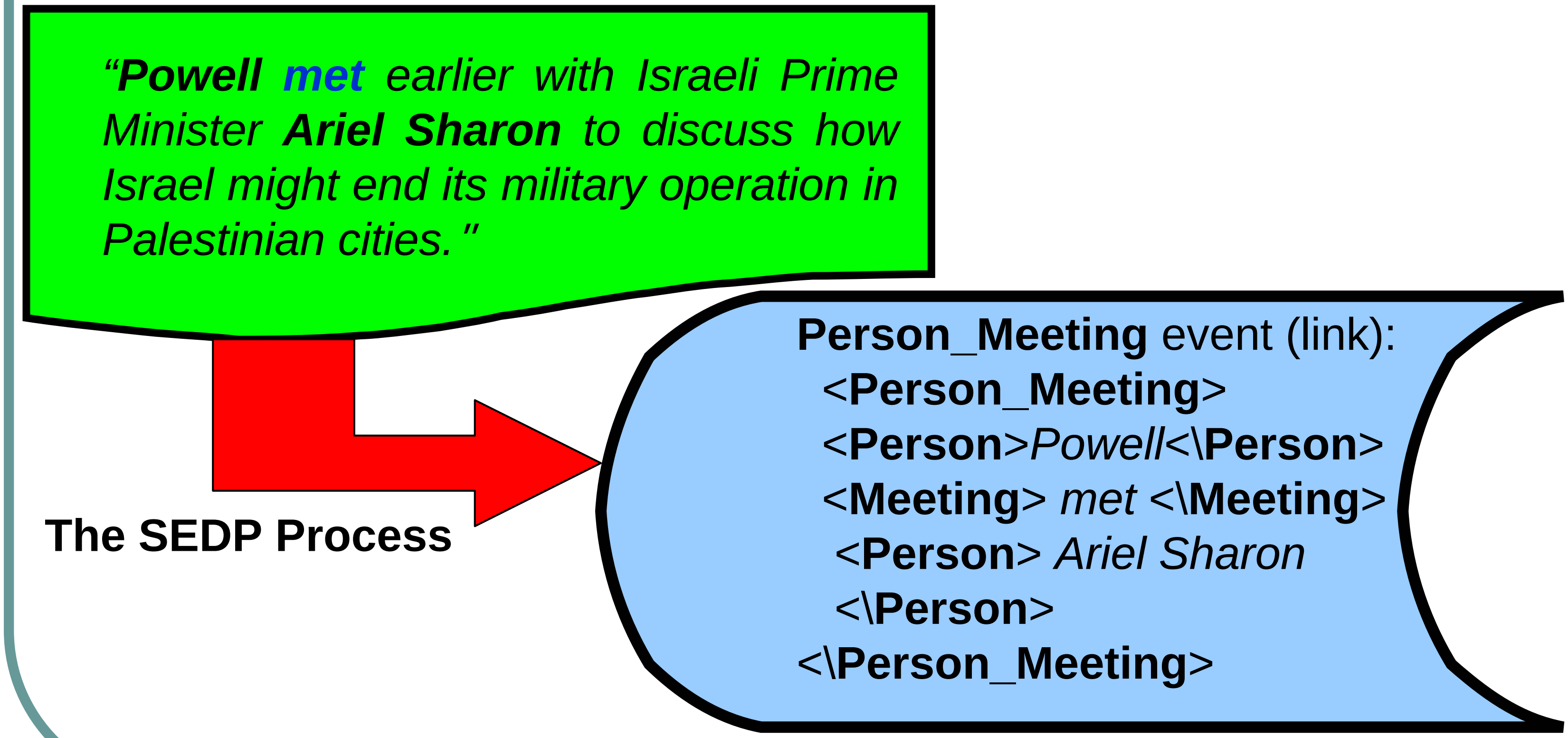
We used the ClearForest® tools SEDP (Semantic Extraction Discovery Process) for creating the semantic links.

The SEDP found **Person_Meeting** events (links).

The experiments and the results

The experiment (cont.)

An example of **Person_Meeting** event (link):



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

The SEDP Process

Person_Meeting event (link):
<**Person_Meeting**>
<**Person**>*Powell*<**Person**>
<**Meeting**> *met* <**Meeting**>
<**Person**> *Ariel Sharon*
<**Person**>
<**Person_Meeting**>

The experiments and the results

The experiment (cont.)

3. **The Semantic links** – The links were created by using noun phrase and verb identification and linguistic and semantic constraints.

We used the ClearForest tools SEDP (Semantic Extraction Discovery Process) for creating the semantic links.

The SEDP found **Person_Meeting** events (links).

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 experiments and the results

The experiment (cont.)

4. **The Precision** – was calculated as the number of correct links (i.e. links that report an actual meeting) divided by the Total number of links found. (8/9)

5. **The Recall** - was calculated as the number of correct links that were found divided by the total number of correct links on the on the whole database (founded at the Preliminary stage). (8/22)

6. We did the same process for Q2 and Q3.

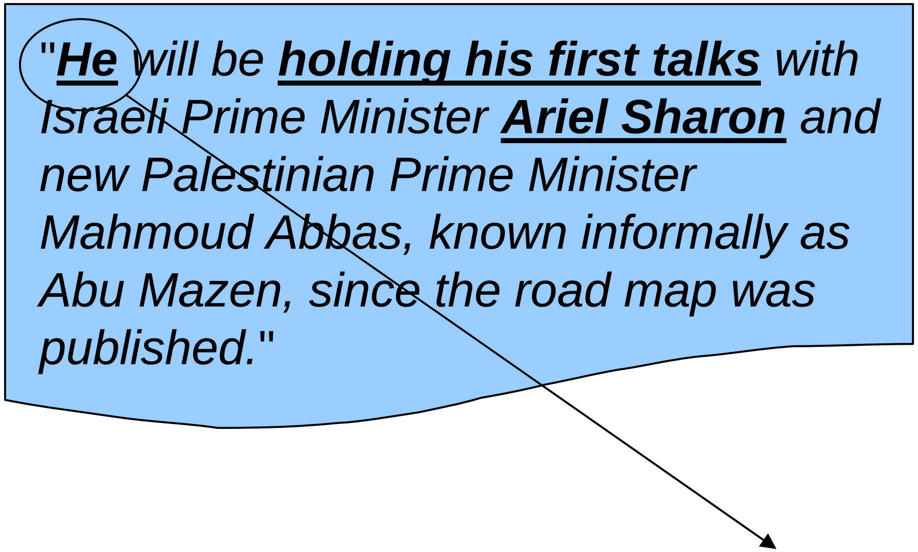
The experiments and the results

The results

	Q1		Q2		Q3	
	Co-occurrence links	Semantic links	Co-occurrence links	Semantic links	Co-occurrence links	Semantic links
Correct links	**20**	**8**	**14**	**5**	**8**	**5**
Total Correct links	**22**	**22**	**15**	**15**	**11**	**11**
Total links	**71**	**9**	**94**	**6**	**9**	**5**
Precision	**28.17%**	**88.89%**	**14.89%**	**83.33%**	**88.89%**	**100%**
Recall	**90.91%**	**36.36%**	**93.33%**	**33.33%**	**72.73%**	**45.45%**

The experiments and the results

The Anaphora problem



*"**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 Colin Powel but the co-occurrence process couldn't find it

The presentation framework

The reasons for combining text-mining and link-analysis

Two links extraction approaches

The experiments and the results

Discussion and Conclusions

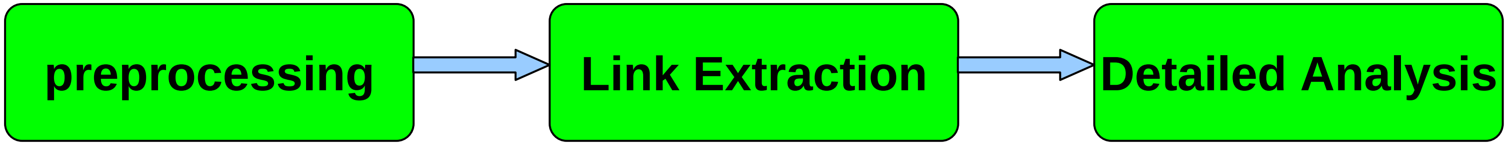
Discussion and Conclusions

- 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.

Discussion and Conclusions

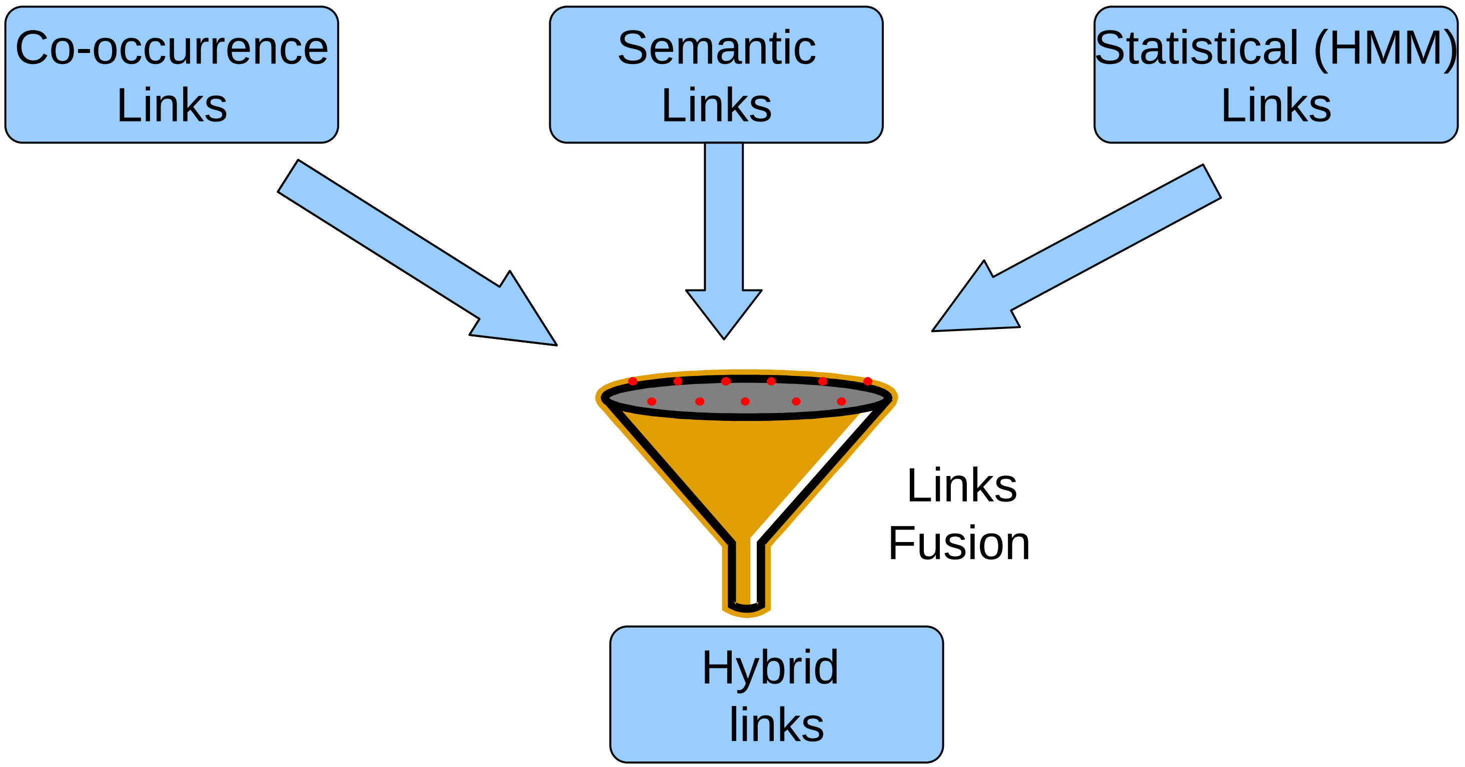
The Knowledge Discovery Process



Co-occurrence links	quick	Time consuming
Semantic links	Time consuming	quick

Discussion and Conclusions

Future plans



Co-occurrence Links

Semantic Links

Statistical (HMM) Links

Links Fusion

Hybrid links

Thank You

Questions ?