Presentation Layout

- Problem and Motivation
- Summary of Contribution
- Defining a Measure of Semantic Relatedness
  - SR Definition
  - OMIOTIS Definition
- Bibliographical Data Classification
- Bibliographical Data Clustering
- Identification of Related Scientific Communities
- Conclusions and Future Work
Problem and Motivation

- In several text related applications (text retrieval, classification, paraphrasing), exact keyword matching misses much information.
  - **Example 1** – Paraphrase Detection
    - Sentence 1: “The shares of the company dropped 14 cents”
    - Sentence 2: “The organization’s stock slumped 14 cents”
  - Sentence 1 is a paraphrase of sentence 2.
    - *share* is synonym to *stock*
    - *drop* is synonym to *slump*
    - *organization* is semantically related to *company*

- In bibliographical data, different terminology used creates even greater problems.
  - **Example 2** – Web Search
    - Query: “search engine log analysis”
    - Semantically related queries: “study of web transactions”, “web queries log analysis”, etc.

Our Solution in a Nutshell

- Unless stated otherwise, we use **Stanford Tagger** for POS tagging, **Porter Stemmer** for stemming, and **TF-IDF** for term weighting.

WordNet 2.0

Pre-processing: Part of speech tagging, phrase detection, stemming, term weighting

Word Sense Disambiguation

Semantic Representation & & Relatedness Computation
Presentation Layout

- Problem and Motivation
- Summary of Contribution
- Defining a Measure of Semantic Relatedness
  - SR Definition
  - OMIOTIS Definition
- Bibliographical Data Classification
- Bibliographical Data Clustering
- Identification of Related Scientific Communities
- Conclusions and Future Work

Summary of Contribution

- A novel measure of semantic relatedness between text segments
- Embedding into bibliographical data classification and clustering
- Empirical evaluation shows clear improvement over traditional term matching techniques
- Novel implementation of the Omiotis semantic relatedness measure
  - all WordNet pair-wise synset relatedness values indexed in a database (11 billion combinations, 600 GB of data, ~1 sec for retrieving 100 term-pair relatedness values)
Presentation Layout

- Problem and Motivation
- Summary of Contribution
- Defining a Measure of Semantic Relatedness
  - SR Definition
  - OMIOTIS Definition
- Bibliographical Data Classification
- Bibliographical Data Clustering
- Identification of Related Scientific Communities
- Conclusions and Future Work

OMIOTIS: A Thesaurus-based measure of Semantic Relatedness

- OMIOTIS is a dictionary-based measure of semantic relatedness.

- It does not require any type of training. It relies on the use of WordNet.

- For the first time, the following important factors are considered in tandem:
  - Semantic path length
  - Depth of senses comprising the path
  - Importance of the semantic edge types
  - All of the available semantic information by WordNet is considered
SR: A Measure of Semantic Relatedness

For all paths between $s_i$, $s_j$ we compute the product $SMC \cdot SPE$, and we keep the maximum value found

$$SR(s_i, s_j) = \max(SCM(s_i, s_j) \cdot SPE(s_i, s_j))$$

Given two terms $t_i$, $t_j$ we compute SR values for all their sense combinations, and we keep the maximum found

$$SR(t_i, t_j) = \max(SR(s_i, s_j))$$

We solve this problem with an altered Dijkstra algorithms and Fibonacci heaps.
OMIOTIS

\[ \text{SR}(t_i, t_{21}) \]

\[ \text{SR}(t_i, t_{22}) \]

\[ \ldots \ldots \]

\[ \text{SR}(t_i, t_{2m}) \]

\[ x(i) = \arg \max_{j\in[1,B]} (\lambda_{i,j} \cdot \text{SR}(t_i, t_j)) \]

\[ \lambda_{i,j} = \frac{2 \text{TFIDF}(t_i, A) \text{TFIDF}(t_j, B)}{\text{TFIDF}(t_i, A) + \text{TFIDF}(t_j, B)} \]

\[ \zeta_1(A, B) = \frac{1}{|A|} \cdot (\sum_{i=1}^{|A|} \lambda_{i,x(i)} \cdot \text{SR}(t_i, t_{x(i)})) \]

\[ \text{OMIOTIS}(A, B) = \frac{1}{2} [\zeta_1(A, B) + \zeta_2(A, B)] \]

Implementation and Complexity

- Index all pair-wise synset SR values in Microsoft SQL Server 2005 (11 billion)
- Dijkstra with Fibonacci heaps
- One-time cost. A DB of 600 GB created.
- Index all term-to-term SR values we meet
- Processing 100 term pairs takes approximately 1 second!

<table>
<thead>
<tr>
<th>Synsets</th>
<th>Edges</th>
<th>Con. Synset Pairs</th>
<th>Avg In-Degree</th>
<th>Avg Out-Degree</th>
<th>Avg Fan-In</th>
<th>Avg Fan-Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>110,490</td>
<td>324,268</td>
<td>11,182,324,723</td>
<td>2.9933</td>
<td>2.9535</td>
<td>103,192.32</td>
<td>101,822.56</td>
</tr>
</tbody>
</table>

- Demo available at: [http://omiotis.hua.gr](http://omiotis.hua.gr)
Presentation Layout

- Problem and Motivation
- Summary of Contribution
- Defining a Measure of Semantic Relatedness
  - SR Definition
  - OMIOTIS Definition
- Bibliographical Data Classification
- Bibliographical Data Clustering
- Identification of Related Scientific Communities
- Conclusions and Future Work

DBLP Bibliographical Data Set

- Parsing of 7 conferences DBLP entries, for years 2006, 2007, and 2008
- Selected to cover various disciplines with potential interest overlap

<table>
<thead>
<tr>
<th></th>
<th>ECDL</th>
<th>ECML/PKDD</th>
<th>FOCS</th>
<th>SIGMOD</th>
<th>VLDB</th>
<th>SODA</th>
<th>KDD</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (2006)</td>
<td>69</td>
<td>149</td>
<td>71</td>
<td>49</td>
<td>136</td>
<td>136</td>
<td>126</td>
<td>786</td>
</tr>
<tr>
<td>Testing (2007 &amp; 2008)</td>
<td>137</td>
<td>255</td>
<td>148</td>
<td>269</td>
<td>161</td>
<td>277</td>
<td>218</td>
<td>1,495</td>
</tr>
</tbody>
</table>

- k-nn used as classifier
- Comparison against
  - k-nn with VSM model and cosine
  - SVM with linear kernel
  - Naive Bayes
Presentation Layout

- Problem and Motivation
- Summary of Contribution
- Defining a Measure of Semantic Relatedness
  - SR Definition
  - OMIOTIS Definition
- Bibliographical Data Classification
- Bibliographical Data Clustering
- Identification of Related Scientific Communities
- Conclusions and Future Work
Bibliographical Data Clustering

- Computed the table with all the pair-wise similarities between paper titles.
- Omi (all edges included in the graph), Omi2 and Omi3 (some edges pruned according to small thresholds) and Cos.
- Used rb graph clustering from the CLUTO suite
- Also compared against standard k-means with cosine

<table>
<thead>
<tr>
<th>Similarity Table</th>
<th>Omi(rb)</th>
<th>Omi2(rb)</th>
<th>Omi3(rb)</th>
<th>Cos(rb)</th>
<th>Cos(k-means)</th>
</tr>
</thead>
<tbody>
<tr>
<td>macro F1 Score</td>
<td>0.622</td>
<td>0.62</td>
<td>0.61</td>
<td>0.611</td>
<td>0.581‡</td>
</tr>
</tbody>
</table>

Presentation Layout

- Problem and Motivation
- Summary of Contribution
- Defining a Measure of Semantic Relatedness
  - SR Definition
  - OMIOTIS Definition
- Bibliographical Data Classification
- Bibliographical Data Clustering
- Identification of Related Scientific Communities
- Conclusions and Future Work
Identifying Related Scientific Communities

- Clustered researchers of two teams (INRIA, Max-Planck) into five groups (Hypergraph partitioning offered by hMetis suite)
- Omiotis groups together researchers from both teams, based on the semantic relevance of their works

Analysis based on co-authorship only

| Cl. 1 | Gagliardi, Galland, Simon, Sais, Chatalic, Roussel, Pernelle, Reynaud, Goasdun, Vento, Safar, Hamdi |
| Cl. 2 | Spaniol, Angelova, Qu, de Melo, Nakashole, Li, Pruski |
| Cl. 3 | Mazeika, Karmici, Elbassioni, Deney, Ramanath, Dagne, Kharlamov, Armant, Ye |
| Cl. 4 | Dietz, Manelescu, Preda, Bourhis, Marinoin, Mrabet, Zoupanos |
| Cl. 5 | Kacimi, Neumann, Theobald, Schenkel, Barberich, Parreira, Cecelius, Pun, Brochart, Sozio, Wang |

Analysis based on Omiotis only

| Cl. 1 | Sozio, Roussel, Sais, Pernelle, Reynaud, Chatalic, Simon, Gagliardi, Goasdun, Vento, Safar, Hamdi |
| Cl. 2 | Dietz, Karmici, Elbassioni, Ou, Ramanath, Kharlamov, Armant |
| Cl. 3 | Mazeika, de Melo, Nakashole, Wang, Bourhis, Marinoin, Pruski, Galland |
| Cl. 4 | Neumann, Theobald, Schenkel, Barberich, Pun, Brochart, Manelescu, Dagne, Preda, Zoupanos, Ye |
| Cl. 5 | Kacimi, Spaniol, Parreira, Cecelius, Angelova, Deney, Li, Mrabet |

Max-Planck researchers in italics, GEMO researchers in Bold

Presentation Layout

- Problem and Motivation
- Summary of Contribution
- Defining a Measure of Semantic Relatedness
  - SR Definition
  - OMIOTIS Definition
- Bibliographical Data Classification
- Bibliographical Data Clustering
- Identification of Related Scientific Communities
- Conclusions and Future Work
Conclusions and Future Work

- Semantic information from word thesauri, like WordNet, can improve the quality of results in bibliographical data mining
- A prototype implementation of the proposed measure, allows for its incorporation in large-scale applications
- For the first time, all pair-wise WN synsets similarities are indexed
- Future Work:
  - Combine distributional similarity
  - Combine knowledge bases (e.g., Wikipedia, and WordNet, like in the case of YAGO)
  - Incorporate more properties on the graph clustering (e.g., co-citation analysis)

Questions

Thank you very much for your attention!

Questions/Comments?