Semantic Relatedness Hits Bibliographical Data

WIDM'09
November 2, 2009, Hong Kong

George Tsatsaronis,
Department of Computer and Information Science
Norwegian University of Science and Technology (**NTNU**)
e-mail: gbt@aueb.gr

Iraklis Varlamis,
Department of Informatics and Telematics
Harokopio University Athens (**HUA**)
e-mail: varlamis@hua.gr

Sofia Stamou,
Computer Engineering and Informatics Department
University of Patras
e-mail: stamou@ceid.upatras.gr

Kjetil Nørvåg,
Department of Computer and Information Science
Norwegian University of Science and Technology (**NTNU**)
e-mail: Kjetil.Norvag@idi.ntnu.no

Michalis Vazirgiannis,
Department of Informatics
Athens University of Economics and Business (**AUEB**)
e-mail: mvazirg@aueb.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

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

The shares of the company dropped 14 cents.

The organization’s stock slumped 14 cents.

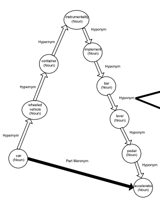
Pre-processing: Part of speech tagging, phrase detection, stemming, term weighting

Word Sense Disambiguation

Semantic Representation && Relatedness Computation

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

WordNet 2.0



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)

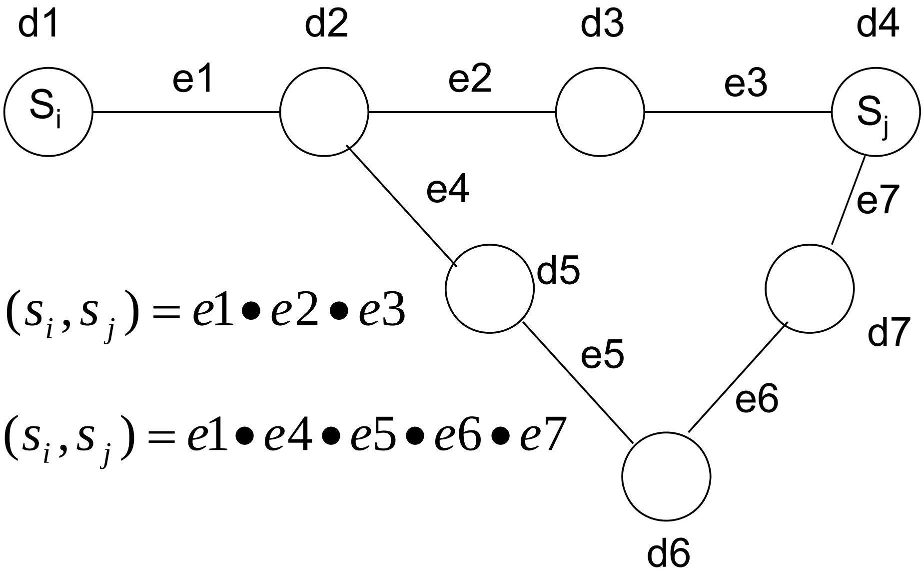
Presentation Layout

- 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 in 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



$$SCM(s_i, s_j) = e1 \bullet e2 \bullet e3$$

$$SCM(s_i, s_j) = e1 \bullet e4 \bullet e5 \bullet e6 \bullet e7$$

$$SPE(s_i, s_j) = \frac{1}{d\max}\frac{2d1d2}{d1+d2} \bullet \frac{1}{d\max}\frac{2d2d3}{d2+d3} \bullet \frac{1}{d\max}\frac{2d3d4}{d3+d4}$$

SR Definition

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

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

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

A

t_{11}
t_{12}
….
t_i
….
t_{1n}

$SR(t_i,t_{21})$
$SR(t_i,t_{22})$
……..
$SR(t_i,t_{2m})$

B

t_{21}
t_{22}
….
t_j
….
t_{2m}

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

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

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

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

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!

Synsets	Edges	Con. Synset Pairs	Avg In-Degree	Avg Out-Degree	Avg Fan-In	Avg Fan-Out
110,490	324,268	11,182,324,723	2.9933	2.9535	103,192.32	101,822.56

- Demo available at: 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

	ECDL	ECML/PKDD	FOCS	SIGMOD	VLDB	SODA	KDD	Total
Training (2006)	69	149	71	99	136	136	126	786
Testing (2007 & 2008)	137	255	148	269	161	277	248	1, 495

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

Classification Results

	7 Conferences Data Set								3 Conferences Data Set							
	Cosine				Omiotis				VSM Cosine				Omiotis			
	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
k=1	0,23§	0,396	0,197‡	0,263§	0,419	0,431	0,411	0,421	**0,576**§	0,752	**0,475**	**0,582**‡	0,75	0,762	0,727	0,743
k=3	0,237§	0,398	0,203‡	**0,269**§	0,408	0,456	0,397	0,425	0,519§	0,706	0,397†	0,508‡	**0,767**	0,805	**0,729**	**0,765**
k=5	0,214§	0,376	0,178§	0,242§	0,408	0,432	0,405	0,418	0,506§	0,697	0,38†	0,492‡	0,75	0,794	0,705	0,7461
k=10	0,105§	0,387	0,082§	0,136§	0,428	0,448	0,42	0,434	0,544§	0,736	0,43†	0,543‡	0,756	0,813	0,701	0,757
k=15	0,092§	**0,4**	0,072§	0,122§	0,441	0,469	0,427	0,447	0,541§	**0,836**	0,423†	0,561‡	0,713	0,8	0,653	0,719
k=20	0,135§	0,384	0,104§	0,164§	**0,444**	0,464	**0,429**	0,445	0,485§	0,826	0,35†	0,492‡	0,693	0,78	0,627	0,696
k=25	0,156§	0,361	0,116§	0,176§	0,439	0,456	0,425	0,441	0,472§	0,157‡	0,333†	0,214‡	0,691	0,803	0,62	0,699
k=30	0,161§	0,356	0,12§	0,18§	0,443	0,479	0,425	0,451	0,472§	0,157‡	0,333†	0,214‡	0,663	0,771	0,588	0,666
k=40	0,281§	0,269†	0,216‡	0,24‡	0,441	**0,488**	0,422	**0,452**	0,472§	0,157‡	0,333†	0,214‡	0,637	0,787	0,552	0,649
k=50	**0,287**§	0,139‡	**0,218**‡	0,169†	0,43	0,474	0,406	0,438	0,472§	0,157‡	0,333†	0,214‡	0,633	0,812	0,545	0,652
k=60	0,264§	0,124§	0,2†	0,152‡	0,429	**0,488**	0,406	0,443	0,472§	0,157‡	0,333†	0,214‡	0,615	**0,824**	0,519	0,637

7 Conferences Data Set								3 Conferences Data Set							
Support Vector Machines				Naive Bayes				Support Vector Machines				Naive Bayes			
A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
0,406‡	0,462	0,401	0,429	0,366§	0,372†	0,39	0,381	0,687§	0,804	0,615†	0,697‡	0,694§	0,737	0,709	0,722

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 prunned according to small thresholds) and Cos.
- Used rb graph clustering from the CLUTO suite
- Also compared against standard k-means with cosine

Similarity Table	**Omi(rb)**	**Omi2(rb)**	**Omi3(rb)**	**Cos(rb)**	**Cos(k-means)**
macro F1 Score	**0.622**	0.62	0.61	0.611	0.581‡

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, Rousset, Pernelle, Reynaud, Goasdoue, Ventos, Safar, Hamdi**
Cl. 2	*Spaniol, Angelova, Qu, de Melo, Nakashole,* **Li, Pruski**
Cl. 3	*Mazeika, Kasneci, Elbassuoni, Denev, Ramanath,* **Dague, Kharlamov, Armant, Ye**
Cl. 4	*Dietz,* **Manolescu, Preda, Bourhis, Marinoiu, Mrabet, Zoupanos**
Cl. 5	*Kacimi, Neumann, Theobald, Schenkel, Berberich, Parreira, Crecelius, Pan, Broschart, Sozio, Wang*

Analysis based on Omiotis only

Cl. 1	*Sozio,* **Rousset, Sais, Pernelle, Reynaud, Chatalic, Simon, Gagliardi, Goasdoue, Ventos, Safar, Hamdi**
Cl. 2	*Dietz, Kasneci, Elbassuoni, Qu, Ramanath,* **Kharlamov, Armant**
Cl. 3	*Mazeika, de Melo, Nakashole, Wang,* **Bourhis, Marinoiu, Pruski, Galland**
Cl. 4	*Neumann, Theobald, Schenkel, Berberich, Pan, Broschart,* **Manolescu, Dague, Preda, Zoupanos, Ye**
Cl. 5	Kacimi, Spaniol, Parreira, Crecelius, Angelova, Denev, ***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?