Semantic Relatedness Hits Bibliographic Data

George Tsatsaronis
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
gbt@idi.ntnu.no

Iraklis Varlamis
Department of Informatics and Telematics
Harokopio University of Athens
Athens, Greece
varlamis@hua.gr

Sofia Stamou
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
stamou@ceid.upatras.gr

Kjetil Nørvåg
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
Kjetil.Norvag@idi.ntnu.no

Michalis Vazirgiannis
Department of Informatics
Athens University of Economics and Business
Athens, Greece
mvazirg@aueb.gr

ABSTRACT

Up-to-date, the effectiveness of bibliographic data retrieval is tightly bound to the performance of traditional keyword matching techniques that require a perfect match between the user-typed query and the indexed keywords of the available data sources. In cases when user requests pursue the retrieval of scientific publications which deal with relevant issues but use different terminology to address them, these techniques are ineffective. In this paper we introduce a novel approach for the thematic organization of bibliographic records that builds upon a semantic relatedness measure we have implemented for this task. In particular, we introduce the Omiotis measure, which captures the semantic relatedness between text segments and enables the thematic organization of the bibliographic data stored in online databases. The experimental evaluation of our measure demonstrates that it can significantly improve the performance of several data mining tasks, such as publications' classification and clustering, compared to existing approaches; even when considering a limited amount of information, i.e., the paper titles.

1. INTRODUCTION

The level of data accessibility in online bibliographic databases is often limited to the metadata descriptors of the publications' contents, e.g., titles, author(s), venues, and year of publication. Consequently, the scientific publications retrieved for some query are those that contain the query terms in the designated metadata elements. Frequently, the information need of users is not always very precise, such as to retrieve a specific paper but might as well be to retrieve publications on a specific subject. For example, the response to a user's request on *search engine logs analysis* will be papers that containing in their title or content one or more of the query terms. As a consequence, publications addressing the subject of *search engine logs analysis* but with a different terminology (e.g. *study of web transactions*) will be omitted from the results.

One way to overcome the above limitation is to equip bibliographic databases with semantics-aware data processing tools that would support the thematic organization of the bibliographic records and would enable the semantic retrieval of scientific publications. In this paper, we introduce a novel approach for organizing publications in bibliographic databases, which relies upon a semantic relatedness measure, named Omiotis. Our relatedness measure quantifies the degree to which different text extracts relate to each other in terms of both lexical and semantic information. In particular, Omiotis estimates the semantic relatedness between publication titles based on the combination of the following: (i) the importance of terms in the publications' titles, (ii) the semantics of title terms as these are determined via WordNet [3], and (iii) the semantic relatedness between the senses identified for the title terms. The advantage of our bibliographic data organization method is that it captures the semantic relatedness of publication titles, even when they use different terminology, without the need for training.

Although, dictionary-based semantic relatedness methods have not been designed to handle such tasks, it is -to the best of our knowledge- the first time that such an approach is presented as an application to the bibliographic data domain, and the results are encouraging. The exploitation of additional semantics afforded by scholarly publication: date, institutional affiliation, departments, citations, co-citation, co-authorship, etc. can further improve search capabilities thus allowing the users of bibliographic database to retrieve information that relates to their search pursuits, rather than simply containing their search query terms.

The following section contains a brief overview of related works. Section 3, explains the basic functionality of Omiotis. Section 4, describes the evaluation process of our measure and presents the results. Finally, Section 5 concludes the paper.

2. RELATED WORK

Bibliographic data organization has attracted the interest of many researchers who study ways of organizing scientific publications in terms of their thematic coverage. Under this scope, they employ

WIDM'09, November 2, 2009, Hong Kong, China.

standard text classification techniques, e.g., Bayesian or Support Vector Machines (SVM) [1], or even Concept Base Vector Space Model [9] in order to assign research papers into appropriate categories. The combined utilization of metadata and full-text information for classifying bibliographic records into appropriate subject classes has been proposed in [7]. However, none of the aforementioned techniques takes into account semantic information from dictionaries or word thesauri.

In the case of scientific community mining from publication records the challenge is to the discovery research communities that share common interests. In [8] a method is proposed that relies on the scientists' publication records in order to create scientific communities. Moreover, in [5] and [11] community mining systems were proposed, which use bibliographic data in order to discover and visualize communities of researchers. Our work is complementary to the above efforts in that our semantic relatedness measure can be explored for capturing the relevance between community member interests as expressed by their publications.

Most of the existing approaches, focus on the feature selection aspect for the sentences classification and pay less attention to the impact that word semantics have on the sentence classification performance. Our dictionary-based method is unsupervised, thus obviates the need for collecting training data samples and its performance does not depend on the quality of the training examples.

3. SEMANTIC RELATEDNESS OF BIBLIOGRAPHIC RECORDS

In this section we introduce the core element of Omiotis, the SR measure [10], which is a measure of semantic relatedness between concepts of a word thesaurus, and proceed with Omiotis, which extends SR to capture the semantic relatedness between text fragments. In the remainder of this paper we distinguish between semantic similarity and semantic relatedness since the former considers only the hierarchical relations (i.e., hypernyms/hyponyms), while the latter takes into account every non-hierarchical semantic relation.

3.1 Background on Semantic Relatedness

SR [10] is a thesaurus-based measure that explores all the different types of semantic links that connect concepts in a thesaurus, in order to estimate the degree to which word pairs semantically relate to each other. WordNet [3], the backbone resource of SR, groups terms that represent a common concept into synonym sets (i.e., synsets) and structures them into a conceptual graph whose nodes represent concepts and edges represent semantic relations between concepts. The SR measure firstly builds the concepts' semantic network, i.e., all the paths that connect the concepts examined, and thereafter computes the weights of these paths by considering: (a) the length of the semantic path; (b) the intermediate nodes' specificity, denoted by the node depth in the thesaurus' hierarchy; and (c) the weight of the semantic edges that compose the path, which depends to the edge type and is analogous to the edge type's frequency of occurrence in the thesaurus O. Eventually, the semantic relatedness for a pair of concepts corresponds to the maximum path weight. The estimated SR values for concept pairs are expanded to their corresponding terms, based on the following definition:

DEFINITION 1. *Let a word thesaurus O, a pair of terms $T = (t_1, t_2)$ for which there are entries in O, S_1 the set of senses of t_1 and S_2 the set of senses of t_2 in O. If S_k, is the set of all senses pairs (s_i, s_j), with $s_i \in S_1$ and $s_j \in S_2$, then the semantic relatedness of T (SR(T,S,O)) is defined as $max\{SCM(S_k, O) \cdot SPE(S_k, O)\}$, for all $k = 1..|S_1| \cdot |S_2|$, where SCM is the accumulative weight of the edges comprising the path and SPE is the accumulative depth of the nodes comprising the path. Semantic relatedness between two terms t_1, t_2 where $t_1 \equiv t_2 \equiv t$ and $t \notin O$ is defined as 1. Semantic relatedness between t_1, t_2 when $t_1 \in O$ and $t_2 \notin O$, or vice versa, is considered 0.*

Motivated by the success of SR towards measuring the semantic relatedness between terms [10], we decided to utilize it as the core measure for quantifying the semantic relevance between sentences (i.e. paper titles in our work). Among all other measures in the bibliography ([2],[4]) SR is the only measure that examines all types of semantic relations within and across Part-of-Speech, has a considerably improved performance compared to existing methods and entails low complexity.

3.2 Identifying Semantically Related Texts

To quantify the degree to which publication titles semantically relate to each other, we build upon the SR measure, which we significantly extend in order to account not only for the terms' semantic relatedness but also for their lexical similarity. This is because publication titles may contain overly-specialized terms (e.g., an algorithm's name) that are inadequately (if at all) represented in WordNet. We begin with the estimation of the terms' importance weights as these are determined by the standard TF-IDF weighting scheme. Thereafter, we estimate the lexical similarity, denoted as $\lambda_{i,j}$ between terms a_i (i.e. the i^{th} term in title A) and b_j (i.e. the j^{th} term in title B) based on the harmonic mean of the respective terms' TF-IDF values, given by:

$$\lambda_{i,j} = \frac{2 \cdot TF\_IDF(a_i, A) \cdot TF\_IDF(b_j, B)}{TF\_IDF(a_i, A) + TF\_IDF(b_j, B)} \quad (1)$$

Having computed the lexical similarity between title terms a_i and b_j, we estimate their semantic relatedness, i.e. $SR(a_i, b_j)$. Our next step is to find for every word a_i in title A the corresponding word b_j in title B that maximizes the product of lexical similarity and semantic relatedness values:

$$x(i) = \underset{j \in [1,|B|]}{\arg\max}(\lambda_{i,j} \cdot SR(a_i, b_j)) \quad (2)$$

Where $x(i)$ corresponds to that term in title B, which entails the maximum lexical similarity and semantic relatedness with term a_i from title A. Consequently, we aggregate the lexical and semantic relevance scores for all terms in title A, with reference to their best match in title B denoted as shown in equation 3:

$$\zeta(A, B) = \frac{1}{|A|} \left(\sum_{i=1}^{|A|} \lambda_{i,x(i)} \cdot SR(a_i, b_{x(i)}) \right) \quad (3)$$

We repeat the process for the opposite direction (i.e. from the words of B to the words of A) to cover the cases where the two titles do not have an equal number of terms. Finally, we derive the degree of relevance between titles A and B by combining the values estimated for their terms that entail the maximum lexical and semantic relevance to one another, given by equation 4.

$$Omiotis(A, B) = \frac{[\zeta(A, B) + \zeta(B, A)]}{2} \quad (4)$$

Algorithm 1 summarizes the computation of Omiotis the complexity of which is strongly related to its base measure of semantic relatedness (SR). In order to improve the scalability of Omiotis, we

Algorithm 1 Omiotis(A,B, Sem, Lex)

Require: Two texts A and B, comprising m and n terms each (a_i and b_j are terms from A and B respectively),
a semantic relatedness measure $Sem : SR(a_i, b_j) \rightarrow (0..1)$,
a weighting scheme of term importance in a text Lex : $TF\_IDF(a_i, A) \rightarrow (0..1)$

Ensure: Find the pair of terms that maximizes the product of Sem and Lex values.

Zeta(A,B)

for all terms $a_i \in A$ **do**
  $x_i := 0$
  $sum(A) := 0$
  for all terms $b_j \in B$ **do**
    $\lambda_{i,j} = \frac{2 \cdot Lex(a_i,A) \cdot Lex(b_j,B)}{Lex(a_i,A)+Lex(b_j,B)}$
    if $x_i < \lambda_{i,j} \cdot Sem(a_i, b_j)$ **then**
      $x_i = \lambda_{i,j} \cdot Sem(a_i, b_j)$
    end if
  end for
  $sum(A) := sum(A) + x_i$
end for
return $sum(A)/|A|$

Omiotis(A,B)

return $\frac{Zeta(A,B)+Zeta(B,A)}{2}$

have pre-computed and stored all SR values between every possible pair of WordNet synsets in a RDBMS. This is a one-time computation cost which dramatically decreases the computational complexity of Omiotis, making it scalable and fast.

4. EXPERIMENTAL EVALUATION

For our evaluation, we utilized different datasets harvested from the DBLP bibliographic database and we carried out three experiments; one for each of the previously described tasks. In the first experiment, we incorporate Omiotis into a k-NN classifier and organize a number of publications into their suitable classes, i.e., their publication venues. We compare the accuracy of results of the k-NN classifier that employs Omiotis against k-NN with cosine, a Support Vector Machines classifier with linear kernel, and a Naive Bayesian Classifier. In the second experiment, we employ Omiotis and cosine similarity for building the adjacency matrices of publication titles. We cluster the set of titles into thematic subsets using the CLUTO clustering suite and the matrices as input and compare results In the third experiment, we compare the performance of Omiotis in automatically identifying researchers with common scientific interests against the standard co-authorship-driven model.

4.1 Bibliographic Data Classification

To evaluate the effectiveness of Omiotis in thematically classifying scientific publications, we relied on the titles of all the papers published in ECDL, ECML/PKDD, FOCS, KDD, SIGMOD, SODA and VLDB conferences between 2006 and 2008. The conferences were selected to cover various disciplines with potential interest overlap, thus constituting a difficult classification problem to solve. The k-NN classifier, was trained against the 786 titles published in 2006 and tested against the remaining $1,495$ titles of years 2007 and 2008. In order to examine whether the number of classes and the potential overlap affects the results, we conducted experiments on a subset of the previous set comprising papers from distinct research communities (ECDL, ECML/PKDD and FOCS from 2006 to 2008). We assessed the performance of the classifiers, using accuracy (A), macro-averaging precision (P), macro-averaging recall (R), and the macro-F1 measure (F1) for each experiment separately. Table 1 presents the results for various values of k in the k-NN classifier. The symbols §, ‡, †, in the results of cosine, indicate a statistically significant difference from the respective Omiotis value at confidence levels 0.99, 0.95, 0.90 respectively. Also, the maximum score found for every evaluation measure in both experiments is highlighted. The results presented in Table 1 indicate that the Omiotis measure outperforms cosine in all cases and with statistical significance at high confidence levels in several cases.

Furthermore, we compare k-NN using Omiotis against a Support Vector Machines classifier that uses linear kernel, and a Naive Bayesian Classifier. For both executions, we used the implementations offered by the Weka platform and compared results as before. Results in Table 2 indicate that k-NN with Omiotis matches the performance of these state of the art classifiers, and in many cases outperforms them with statistical significance.

In all, the results of the classification task clearly demonstrate that Omiotis constitutes a good alternative, to traditional techniques in discriminating between thematically related titles. In this respect, Omiotis can definitely be explored as the core measure for semantic classification since it has a stable performance and it is quite effective in identifying the best matching topic for a paper among numerous topics.

4.2 Bibliographic Data Clustering

In this task we use the small dataset (ECDL, ECML/PKDD and FOCS papers between 2006 and 2008) and cluster all the titles into three groups. We employ the CLUTO suite [6] and more specifically the default rb algorithm, a graph based clustering algorithm, and examine in what degree titles from the same publication venue have been clustered together. We use interchangeably Omiotis and Cosine measures for creating the publication titles' adjacency matrix upon, which clustering would operate and compare results. Nevertheless, to ensure that the performance of Omiotis is not biased by the size of the original graph (Omi), we apply different threshold values to the Omiotis scores that prune several edges of the graph and result in simpler graphs. All edges with Omiotis $<10^{-3}$ and <0.045 have been pruned in $Omi2$ and $Omi3$ respectively, thus providing a graph ($Omi3$) of size comparable to that of the Cosine graph (Cos) and a second graph ($Omi2$) with almost half the edges of the Omiotis graph (Omi).

To compare the clustering performance between the cosine and Omiotis, we used the four similarity matrices and all the criteria that rb supports. For our comparisons, we employed F-measure, which computes the degree of correspondence between the predefined categorization of paper titles into conferences and the resulting clustering schemas. Furthermore, we conduct in the same data set an experiment with k-means ($k = 3$), using cosine as the similarity measure and the Weka implementation for the algorithm.

The F-measure values (0.622 for Omi(rb), 0.62 for $Omi2$(rb), 0.61 for $Omi3$(rb), 0.611 for Cos(rb) and $0.581^{\ddagger}$ for Cos(k-means) indicate that Omiotis has an improved overall clustering performance compared to its variations. This implies that our measure can be readily employed in bibliographic data clustering applications without any prior need for fine-tuning its parameters.

4.3 Identifying Scientific Communities

In order to evaluate the effectiveness of Omiotis in automatically discovering researchers with common interests based on their publication records we considered the DBLP entries for the researchers of two distinct research teams, i.e. the GEMO group at INRIA Orsay and the Database and Information Systems group at Max-Planck-Institute for Informatics. In particular, we relied

	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

Table 1: Classification results for cosine and Omiotis. Confidence levels: †=0.90, ‡=0.95, §=0.99

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

Table 2: Classification results for Support Vector Machines and Naive Bayes. Confidence levels: †=0.90, ‡=0.95, §=0.99

on the multi-level k-way Hypergraph Partitioning algorithm of the hMetis suite [6] in order to built a hypergraph with researchers as nodes and hyper-edges that connect more than two nodes at the same time. For building hyper-edges we tried two different implementations, one that relies exclusively on co-authorship relations extracted from DBLP and one that relies on high Omiotis values (above a 0.5 threshold) between the researchers' paper titles. In the former a hyper-edge corresponds to a paper written by some researchers and connects the respective researcher nodes, whereas in the latter a hyper-edge connects all the authors of related titles. The first implementation resulted in 414 hyper-edges and the second to an additional set of 41 (i.e. 455 in total) hyper-edges. We set the number of desired groups to five and cluster the two hypergraphs. The co-authorship-based method gives two "clean" groups that comprise members of one team only and one group that contains mainly GEMO researchers. On the other hand, the Omiotis-based method groups together researchers from both teams based entirely on the semantic relevance between their published works. For example using the Omiotis-based method Neumann, Theobald, Schenkel, Berberich, Pan and Broschart from Max Planck are grouped together with Manolescu, Dague, Preda, Zoupanos and Ye from Gemo, based on their common interests, which relate to web search and XML. Our findings give some preliminary evidence that Omiotis could be fruitfully explored as an alternative method for discovering potential collaborations among researchers.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a novel approach for the thematic organization of the bibliographic database contents. The contribution of our approach lies on the provision of a novel measure for capturing both the semantic and the lexical relevance between small texts (i.e. publication titles in this work) rather than find the optimal clustering and classification schemes for bibliographic data organization. Therefore, we believe that our measure can be fruitfully explored in several other data mining applications and is on next plans to extend the application of our measure to other tasks that involve thematic organization of texts.

6. REFERENCES

[1] R. Angelova and G. Weikum. Graph-based text classification: learn from your neighbors. In *SIGIR*, 2006.

[2] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[3] C. Fellbaum. *WordNet – an electronic lexical database*. MIT Press, 1998.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of the 20th IJCAI*, pages 1606–1611, 2007.

[5] R. Ichise, H. Takeda, and K. Ueyama. Community mining tools using bibliography data. In *Proc. of the 9th Intl. Conf. on Information Visualization*, pages 953–958, 2005.

[6] G. Karypis. The karypis lab homepage. In *http://www.cs.umn.edu/ karypis*.

[7] A. Montejo-Raez, L. Urena-Lopez, and R. Steinberger. Text categorization using bibliographic records: beyond document content. In *Proc. of the 21st Conference of the Spanish Society for NLP*, pages 119–126, 2005.

[8] S. Rodriguez, I. Oliveira, and J. de Souza. Competence mining for virtual scientific community creation. *Intl. Journal of Web Based Communities*, 1(1):90–102, 2004.

[9] T. Shimano and T. Yuakawa. An automated research paper classification method for the ipc system with the concept base. In *Proc. of the NTCIR-7 Workshop Meeting*, 2008.

[10] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis. Word sense disambiguation with semantic networks. In *TSD*, 2008.

[11] O. Zaiane, J. Chen, and R. Coebel. Bdconnect: Mining research community on dblp data. In *Web Mining and Social Network Analysis Workshop, ACM SIGKDD*, 2007.