Combining Link and Content Information for Scientific Topics Discovery

Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles
Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex
{chikhi,rothenburger,aussenac}@irit.fr

Abstract

The analysis of current approaches combining links and contents for scientific topics discovery reveals that the two sources of information (i.e. links and contents) are considered to be heterogeneous. Therefore, in this paper, we propose to integrate link and content information by exploiting the links semantics to enrich the textual content of documents. This idea is then implemented in two variants: a local content enrichment method, and a global content enrichment technique.

Experiments carried out on two real-world datasets show the good performances of our approach over state of the art techniques that combine citation and content information for scientific topics discovery.

1. Introduction

In the past, a scientist was able to master many different scientific disciplines. Nowadays, the situation has changed greatly; a scientist is no longer able to know every detail of, even, his own specialty. Actually, this setback is due to the exponential growing of science where many new journals, conferences and topics are created each year. For instance, it is reported that the NASA ADS (Astrophysics Data System)[1] digital library contains more than 7 million papers with an increase larger than 100.000 papers each year. This amount of knowledge is obviously too huge to be read and digested by a human researcher.

An experienced researcher can easily identify the main topics addressed in a paper with a glance to the title, abstract, and eventually the bibliographic-reference list. Unfortunately, a novice researcher do not have these skills, and may spend a great portion of his/her time in selecting the papers which "would be" relevant to his research topic. Clearly, intelligent tools to help scientists in the exploration of these deep scientific repositories turn out to be more than essential. Such assistance tools may consist, for example, in the *related papers* functionality proposed by some databases such as Citeseer[2]. The success of such utilities depends mainly on their ability to correctly identify the scientific topics addressed in a paper.

Traditionally, two approaches have been considered for the identification of scientific topics. The first one is the result of *bibliometrics,* which focused on the analysis of citations between documents [4]. Basically, citation analysis consists in using some bibliometric similarity measures such as co-citation or bibliographic coupling along with a clustering algorithm. Experiments carried out by many researchers proved the usefulness of citation analysis to find the scientific topics embodied in a collection of documents. These successful results support the idea that bibliographic references do carry some semantic information about the scientific fields organization.

The second approach, which originated from *information science*, focused on the storage, analysis and retrieval of the content of documents [4]. Technologies used by this community include: the vector space model for documents representation, term weighting schemes (e.g. TFIDF), similarity measures (e.g. cosine), and clustering algorithms (e.g. K-means). A more elaborated and well-known technique for text analysis is the Latent Semantic Indexing (LSI) method [8]. Given a term-document matrix, the LSI groups "magically" documents into clusters where each group corresponds to a latent concept (i.e. topic).

More recently, a third approach has emerged which consists of combining the two kinds of information (i.e. links and contents) to find the scientific topics [9][13]. Many approaches have been proposed having their own limitations. The main drawback, which we have noticed, is that hybrid methods consider the two sources of information as heterogeneous. In other words, they use the two information sources without exploiting the semantic relationships between links and contents.

Therefore, in this paper, we propose a new approach which consists of using link information to enrich the textual representation of documents before mining their content. Our approach is then evaluated on two bibliographic datasets, and is shown to be superior to other approaches which combine link and content mining.

[1] http://adswww.harvard.edu
[2] http://citeseer.ist.psu.edu

The idea of enhancing documents representation has been used previously in the context of web page classification. For instance, Yang et al. [23] used the words of linked and linking pages to enhance hypertext classification. They carried out experiments on various web datasets, and concluded that the utility of the neighboring web pages depends strongly on the nature of the dataset. In the same vein, Oh et al. [19] used the web pages in the neighborhood of a webpage and observed that such an approach introduces some noise in the representation of pages, and thus deteriorates the classification performance. While our approach is closely related to these methods, we argue that they are, fundamentally, very different since web hyperlinks and bibliographic citations do not have the same semantics [2].

The rest of the paper is organized as follows. In the next section, we review existing techniques for combined link and content mining. The proposed approach is then described is Section 3. Sections 4 and 5 present the experiments, and Section 6 concludes the paper.

2. Existing approaches to link and content combination for scientific topics discovery

Here, we propose to classify existing approaches to link and content combination into two categories; each one combines the link and content information differently.

2.1. Similarity-based combination

A simple way to combine link and content information is through an integrated similarity matrix computed from the two information sources. The idea of this approach is that, given two similarity matrices between documents based respectively on link and content information, compute a similarity matrix which takes into account the two similarity measures (Figure 1a). Formally, if $\mathbf{S_L}$ is the pairwise similarity matrix based on link information, and $\mathbf{S_C}$ is the pairwise similarity matrix based on content, then a global similarity matrix $\mathbf{S}$ is obtained by: $\mathbf{S} = f\left(\mathbf{S}_L, \mathbf{S}_C\right)$.

The function f can be, for instance, a multiplication of the two similarities [17]: $f\left(\mathbf{S}_L, \mathbf{S}_C\right) = \mathbf{S}_L \cdot \mathbf{S}_C$, or a weighted linear combination [17],[13] : $f\left(\mathbf{S}_L, \mathbf{S}_C\right) = \left(1-\alpha\right)\mathbf{S}_L + \alpha\mathbf{S}_C$ where $0 \le \alpha \le 1$.

As naive as it may appear, the weighted linear combination of link and content similarities was shown by Janssens [13] to be very effective for scientific topics discovery. He shows that the linear combination performs as good as a more elaborate combination approach based on the Fisher's inverse chi-square method.

2.2. Joint-factorization-based combination

Instead of dealing with similarities, another family of techniques uses directly the original representations of data (Figure 1b). In our case, these representations correspond to the links and content views.

In [7], Cohn & Hoffmann proposed PHITS-PLSA, a probabilistic model for both link and content generation. The PHITS-PLSA algorithm consists in a joint factorization of the adjacency matrix and the word-document matrix.

Likewise the weighted linear combination of similarities, PHITS-PLSA also uses a weighting factor. This factor balances the importance given to each source of information. In the two extremes, when $\alpha = 0$ (resp. $\alpha = 1$), the algorithm is equivalent to a text analysis by PLSA [12] (resp. to a link analysis using PHITS [6]).

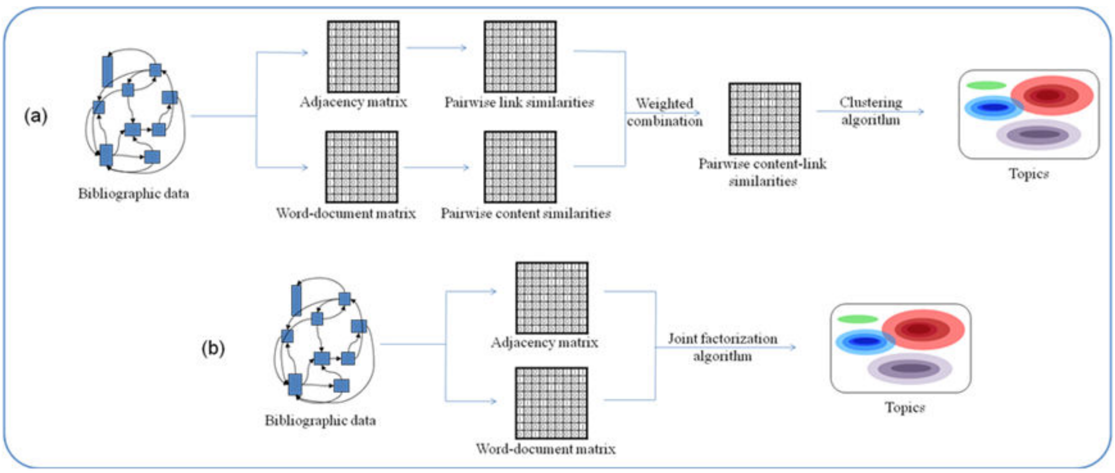

Figure 1 – (a) Similarity-based combination; (b) Joint-factorization-based combination

More recently, Zhu et al. [24] proposed, in the same spirit of PHITS-PLSA, a method for document classification which jointly factorizes link and content matrices. In contrast to PHITS-PLSA, which uses inlink information only, Zhu et al. use both inlink and outlink information. No detail, however, has been given by the authors on how the tradeoff between links and contents is made.

A similar approach was used by Erosheva et al. [9] to find scientific topics. They simply replaced the PLSA model with the more recent LDA model [3]. In our experiments, we have considered only the approach based on PHITS-PLSA. In fact, the Erosheva et al.'s approach, based on LDA, uses some hyper-parameters which are very decisive on the performances of the algorithm. Unfortunately, finding the good values for these hyper-parameters turns out to be a tricky task.

3. Proposed approach

The analysis of current approaches combining links and contents for scientific topics discovery reveals that the two sources of information (i.e. links and contents) are considered to be heterogeneous. Therefore, in this paper, we propose to integrate link and content information by exploiting the links semantics to enrich the textual content of documents (Figure 2).

3.1. Citations semantics

For a long time, many researchers in bibliometrics have examined the *motivations* of authors to cite other papers. For example, Eugene Garfield, founder of the ISI (Institute for Scientific Information), identified the following motivations behind citing [10]: paying homage to pioneers, identifying methodology or equipment, providing background reading, etc.

The main idea of our work is to use citation information as a mean for enhancing the textual representation of documents. More precisely, we view a scientific paper as a small piece of knowledge which cannot be correctly and entirely interpreted if it is taken solely. If taken separately, a scientific paper is much like a concept taken from an ontology without knowing the relationships of this concept to other concepts in the ontology; it is almost impossible to figure out the correct meaning of such an isolated concept.

The words present in a scientific paper are generally not sufficient to fully characterize it because in scientific papers, authors often make some assumptions on the background knowledge of their readers. Let's suppose, for example, an author writing a paper which is based on an old theory. Unfortunately, most of the potential readers of the paper would be unfamiliar with this old theory. Since describing the full details of the theory is outside the scope of the paper (and will take too much space), the author will often simply redirect interested readers to a more detailed reference about the theory in question.

However, citations in a paper are not always useful and necessary for the interpretation of the document. Nevertheless, the majority of citation analysis researchers agree on the *relevance* of most of the cited documents with respect to the citing document (which is not the case of web pages for example) [11].

3.2. Textual content enrichment from the bibliographic context

Here we introduce the notion of *bibliographic context*, which is the core of our textual content enrichment methodology. We denote by bibliographic context the set of all the documents necessary to correctly interpret and characterize the textual content of a document. Virtually, as it is defined, the bibliographic context of a document would correspond to a huge set of documents. Thus, to make our approach feasible, we propose two simple formulations of this notion: the *local* bibliographic context and the *global* bibliographic context.

3.2.1. Content Enrichment from the local bibliographic context. The local bibliographic context *LBC* of a scientific paper *P* is defined as the set of documents which are directly connected to *P*. According to this definition, three cases are possible for the LBC: citing (i.e. inlinks) documents, cited (i.e. outlinks) documents, and both of them.

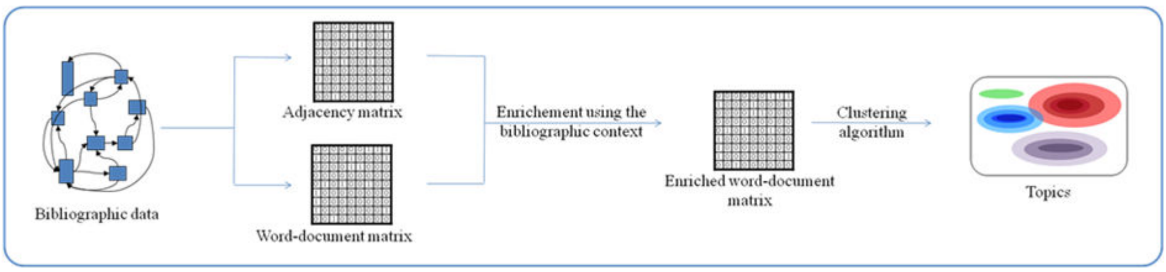

Figure 2 – Content enrichment-based combination of links and contents

Formally, the three kinds of local bibliographic contexts can be expressed as:

$$LBC_I(P) = \{\text{documents } D \text{ s.t. } D \in \text{Inlinks}(P)\}$$

$$LBC_O(P) = \{\text{documents } D \text{ s.t. } D \in \text{Outlinks}(P)\}$$

$$LBC_{IO}(P) = \left\{ \begin{array}{l} \text{documents } D \text{ s.t.} \\ D \in (\text{Inlinks}(P) \cup \text{Outlinks}(P)) \end{array} \right\}$$

Once the bibliographic context of each document is determined, the content enrichment is then performed using the following simple procedure:

For every document $D \in LBC(P)$

$$\mathbf{E}(W,P) = \alpha \mathbf{T}(W,P) + (1-\alpha)\frac{\mathbf{T}(W,D)}{|LBC(P)|}$$

where **T** is the original word-document matrix, **E** is the enriched word-document matrix, and $\alpha \in [0,1]$ controls the importance given to the textual content imported from the bibliographic context. The division by $|LBC(P)|$ aims at normalizing the importance of each document in the bibliographic context. This normalization avoids the content of documents having a large bibliographic context from being understated by the content of their bibliographic context.

Figure 3 shows a toy example of content enrichment from cited documents (i.e. outlinks).

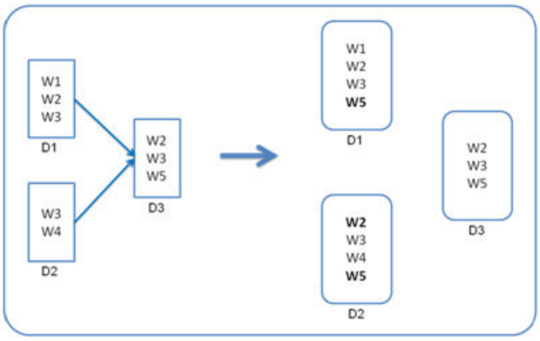

Figure 3 – Content enrichment from cited documents (words in bold denote imported ones)

3.2.1. Content Enrichment from the global bibliographic context. An alternative to using the content of directly neighboring documents is to use the content of documents which are similar to the current document according to the global topography of the citation graph. It is possible for instance to enrich the content of a document from the content of documents with whom it is co-cited. To this end, we propose to use three global similarity indicators namely co-citation [20], bibliographic coupling [14] and the Amsler factor [1]. The global bibliographic contexts according to these three indicators are defined as:

$$GBC_{Co}(P) = \left\{ \begin{array}{l} \text{documents } D \text{ s.t. } \exists Q \text{ where:} \\ P \in \text{Outlinks}(Q) \ \wedge \ D \in \text{Outlinks}(Q) \end{array} \right\}$$

$$GBC_{Bc}(P) = \left\{ \begin{array}{l} \text{documents } D \text{ s.t. } \exists Q \text{ where:} \\ P \in \text{Inlinks}(Q) \ \wedge \ D \in \text{Inlinks}(Q) \end{array} \right\}$$

$$GBC_{Am}(P) = \left\{ \begin{array}{l} \text{documents } D \text{ s.t. } \exists Q \text{ where:} \\ P \in (\text{Inlinks}(Q) \cup \text{Outlinks}(Q)) \\ \wedge\ D \in (\text{Inlinks}(Q) \cup \text{Outlinks}(Q)) \end{array} \right\}$$

Content enrichment from the global bibliographic context is performed using the same procedure as the one of content enrichment from the local bibliographic context.

Figure 4 illustrates the content enrichment approach from the global bibliographic context based on the bibliographic coupling.

We note here that we have deliberately chosen to consider a document to be in the global bibliographic context of another document only if they are, for instance co-cited, at least one time by a third document. Another possibility is to use a threshold *k*, such that a document *p* is considered to be in the global bibliographic context of a document *q* only if they are, for instance co-cited, at least *k* times.

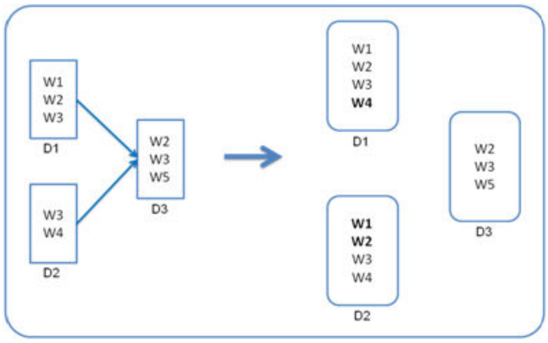

Figure 4 – Content enrichment from bibliographically coupled documents (words in bold denote imported ones)

4. Experimental setup

4.1. Datasets

To evaluate our approach and compare it with other approaches, we have used two datasets of scientific papers. The first dataset is a subset of the Cora collection,

Dataset	*Documents*	*Categories*	*Average words per document*	*Average links per document*	*Documents having inlinks*	*Documents having outlinks*
Cora	2708	7	62	2	1565	2222
Citeseer	2994	5	32	1.43	1760	2099

Table 1 – Datasets and their properties

which is a set of more than 30,000 papers in the computer science field [18]. Our subset consists of 2700 documents where each one belongs to one of the following categories: Neural networks, genetic algorithms, reinforcement learning, learning theory, rule learning, probabilistic learning methods, and case based reasoning.

The second dataset consists in collection of 3000 papers extracted from the Citeseer database. The documents are classified into one of the following topics: Agents, databases, information retrieval, machine learning, and human computer interaction. Statistics on the two datasets are presented in Table 1.

4.2. Evaluation measures

In the machine learning literature, many clustering assessment measures can be found. In our experiments, we have used the F-measure and the normalized mutual information as performance criterions.

The traditional F-measure corresponds to the arithmetic mean between precision and recall. It is computed by the formula:

$$\text{F-measure} = \frac{2 \times precision \times recall}{precision + recall}$$

Because clustering indicators such as the F-measure or even the entropy or the purity are known to be biased by the size of different clusters and classes [21], we use also a second evaluation criterion based on information theory, which is less sensitive to variation in clusterings. The normalized mutual information [21] between two categorizations (i.e. clusterings) A and B is defined as:

$$\text{NMI}(A,B) = \frac{H(A) + H(B) - H(A,B)}{\sqrt{H(A) \cdot H(B)}}$$

where *H(A)*, *H(B)* are respectively the entropy of A and B; *H(A,B)* is the joint entropy of A and B. The factor in the denominator is the normalization factor.

4.3. Clustering algorithm

The different approaches we deal with in this paper need an unsupervised learning algorithm as a final step in the scientific topics discovery process. Therefore, many clustering algorithms can be used such as Hierarchical agglomerative clustering, K-means, Nonnegative Matrix Factorization (NMF), PLSA, etc.

We have chosen to use the NMF [16] algorithm for several reasons. First, it has been empirically proven to be effective for the analysis of text data [22] and link data [5]. Second, NMF is a soft clustering algorithm which allows finding overlapping clusters. It is also able to give the most representative words and documents for each cluster (i.e. for each scientific topic in our case). Last but not least, NMF is simple and efficient. Technically, it consists in applying iteratively two simple update rules.

5. Experimental results

In Figures 5 and 6, we report the experimental results of applying, on the Cora and Citeseer datasets, four scientific topics discovery algorithms, namely: weighted linear combination (WLC), PHITS-PLSI, local content enrichment (LCE), and global content enrichment (GCE). The four algorithms are evaluated in three different contexts: whether inlink, outlink, or inlink+outlink information is used.

On a quantitative basis, the analysis of the obtained results shows that our approach (i.e. content enrichment) significantly outperforms the other methods. In other words, there exists in each case a value of the combination factor (i.e. alpha) for which our approach achieves the best performance.

Qualitatively, several aspects of the obtained results can be noticed. The first one concerns the tricky task of combination factor determination. While in our approach, fixing alpha to 0.5 yields a close to optimal result, in the other approaches (WLC and PHITS-PLSI) the best value for alpha ranges from 0.15 to 0.8; this makes the choice of the alpha value unpredictable. We note also that the performances of WLC and PHITS-PLSI vary greatly depending on the value of alpha. Actually, this variability is due to the combination process in the WLC and PHITS-PLSI algorithms, which merges two information sources different in nature. Indeed, Janssens [13] shows that citations and contents have different statistical distributions. We can also observe from Table 1 that documents have on average much more words than links.

The second aspect is related to the impact of inlinks and/or outlinks on the scientific topics discovery performance. Except when using only outlinks with the Cora dataset (Figures 5c and 5d), the LCE and GCE algorithms appear to have the same performance. Moreover, all the compared techniques achieved best performance when both inlinks and outlinks were used.

Lastly, compared to Figure 6 (i.e. Citeseer results), Figure 5 (i.e. Cora results) shows that the bibliographic context is more valuable in the Cora dataset than it is in the Citeseer dataset. This observation can be explained by the difference in the amount of citations existing in each dataset. Statistics in Table 1 show that the Cora dataset is richer in links than the Citeseer dataset. Hence, the bibliographic context in the Cora dataset will contain more documents than in the Citeseer dataset.

6. Conclusion and future work

In this paper, we have proposed a new approach for scientific topics discovery. Our approach exploits citation information as a mean to enrich the textual representation of documents. This idea has been implemented in two variants. The first one uses the content of directly neighboring documents and was called local content enrichment. The second one uses the content of documents which are "citationally" related to the paper in hand; this approach was named global content enrichment.

Experiments carried out on two real-world datasets have shown the good performances of our approach over state of the art techniques that combine citation and content information for scientific topics discovery. More precisely, the proposed approach proved the utility of words taken from the bibliographic context. Experimental results showed also that both inlinks (i.e. citing documents) and outlinks (i.e. cited documents) are useful for determining the bibliographic context. An additional interesting result concerns the small sensitivity of our approach to the combination factor (i.e. α); the other approaches achieved very poor performances for an inadequate value of the combination factor.

Currently, we are improving our approach by adding a weighting scheme a la TFIDF which determines the importance of the words borrowed from the bibliographic context. It may be useful for instance to exploit the authority or the hubity of a document [15] to determine the importance that should be given to its words.

As our future work, we plan to exploit not only citation and content information, but also other information sources such as author information, conference or journal where the paper has been published, and tag information from current web 2.0 sites such as Bibsonomy[3].

Acknowledgment

This work was supported in part by the INRIA under Grant 200601075.

References

[3] http://www.bibsonomy.org

[1] R. Amsler. Application of Citation-based Automatic Classification. Austin, TX: The University of Texas at Austin, Linguistics Research Center, Internal Technical Report 72-12, 42p., 1972.

[2] L. Bjorneborn. Small-world link structures across an academic web space: A library and information science approach. Ph.D. Thesis, Royal School of Library and Information Science, Denmark, 2004.

[3] D. M. Blei, Ng A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, 2003.

[4] J. P. Carlisle, S. W. Cunningham, A. Nayak, and A. Porter. Related problems of knowledge discovery. In Proc. Of the 32nd Hawaii Intl. Conf. On System Sciences. Hawaii, USA, 1999.

[5] N. F. Chikhi, B. Rothenburger, and N. Aussenac-Gilles. Authoritative documents identification based on nonnegative matrix factorization. In proc. of the IEEE Intl. Conf. on Information Reuse and Integration. Las Vegas, USA, 2008.

[6] D. Cohn, and H. Chang. Learning to probabilistically identify authoritative documents. In Proc. of the 17th ICML, 2000.

[7] D. Cohn, and T. Hofman. The missing link - A probabilistic model of document content and hypertext connectivity. In Proceedings of the 13th NIPS Conference. Vancouver, Canada, pp. 430–436, 2001.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and D. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(7):391-407, 1990.

[9] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences (PNAS), vol. 101, pp. 5220–5227, 2004.

[10] E. Garfield. Can citation indexing be automated? Essay of an Information Scientist, 1:84-90, 1962.

[11] S. P. Harter. Psychological relevance and information science. Journal of the American Society for Information Science, 43(9):602-615, 1992.

[12] T. Hofmann. Probabilistic latent semantic analysis. In Proc. of the 15th UAI Conference, 1999.

[13] F. Janssens. Clustering of scientific fields by integrating text mining and bibliometrics. Ph.D. Thesis, Katholieke Universiteit Leuven, Belgium, 2007.

[14] M. M. Kessler. Bibliographic coupling between scientific papers. American Documentation, 14(1), 10–25, 1963.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, 1999.

[16] D. Lee, and H. Seung. Algorithms for non-negative matrix factorization. In Proc. of NIPS, pages 556–562, 2000.

[17] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic detection of semantic similarity. In Proc. of the WWW 2005 Conf., Chiba, Japan, 2005.

[18] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. Information Retrieval Journal, 3 (200), 127–163, 2000.

[19] H.J. Oh, S. H. Myaeng, and M.H. Lee. A practical hypertext categorization method using links and incrementally available class information. In Proc. of the 23rd annual Intl. ACM SIGIR Conf. Athens, Greece, pp. 264–271, 2000.

[20] H. G. Small. Co-citation in the scientific literature: A new measure of relationship between two documents. Journal of the American Society for Information Science, 24(4), 265–269, 1973.

[21] A. Strehl, Relationship-based clustering and cluster ensembles for high-dimensional data mining. Ph.D. Thesis, Austin University, USA, 2002.

[22] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In Proc. of the 26th annual Intl. ACM SIGIR Conf. Toronto, Canada, pp. 267–273, 2003.

[23] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. Journal of Intelligent Information Systems, 18(2):291-241, 2002.

[24] S. Zhu, Yu K., Y. Chi, and Y. Gong. Combining content and link for classification using matrix factorization. In proc. of the 30th annual intl. ACM SIGIR conf. on Research and development in information retrieval. Amsterdam, The Netherlands, pp. 487–494, 2007.

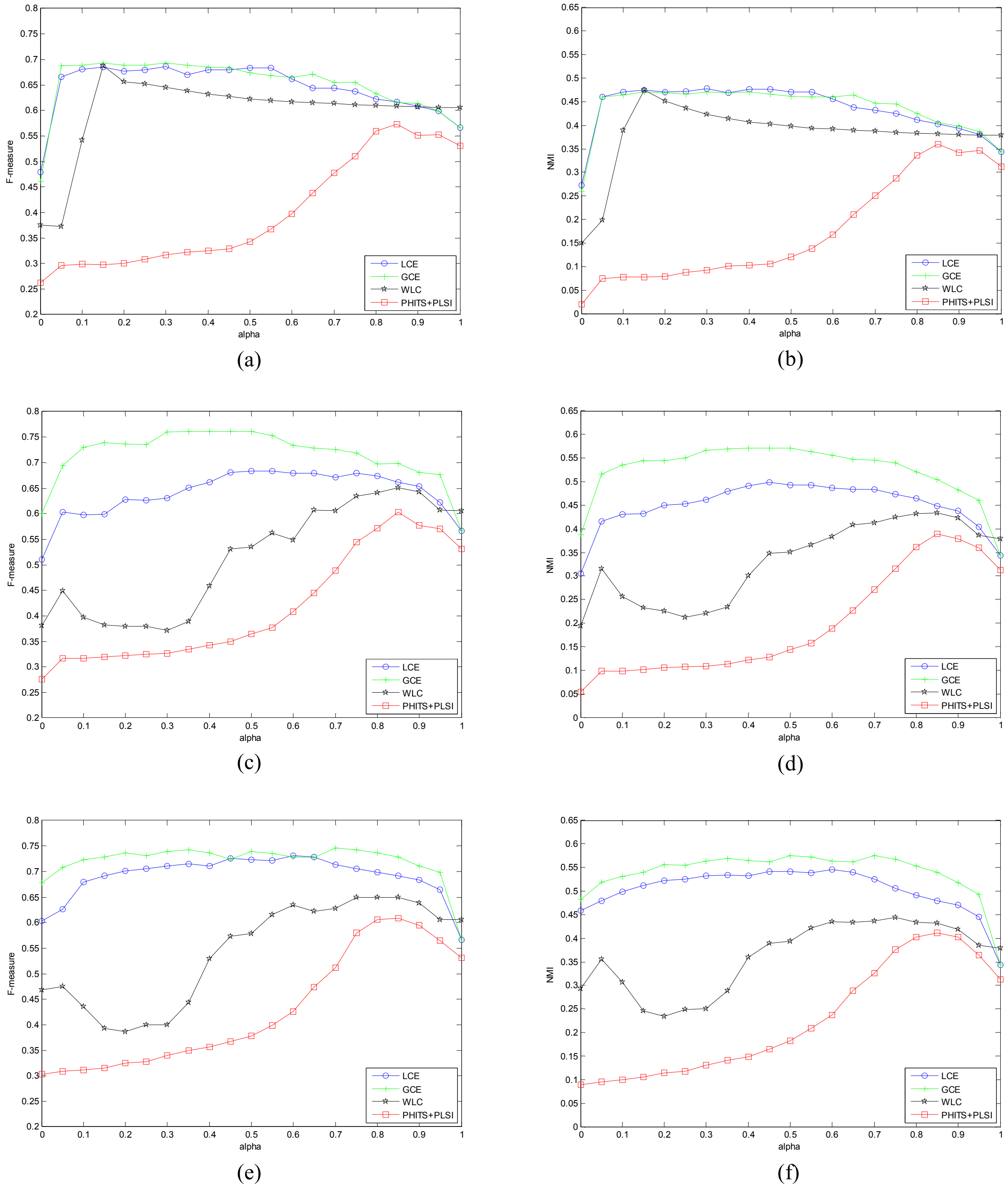

Figure 5 – Results on Cora dataset using inlinks (a-b), outlinks (c-d) and inlinks+outlinks (e-f).
(LCE: Local Content Enrichment; GCE: Global Content Enrichment; WLC: Weighted Linear Combination)

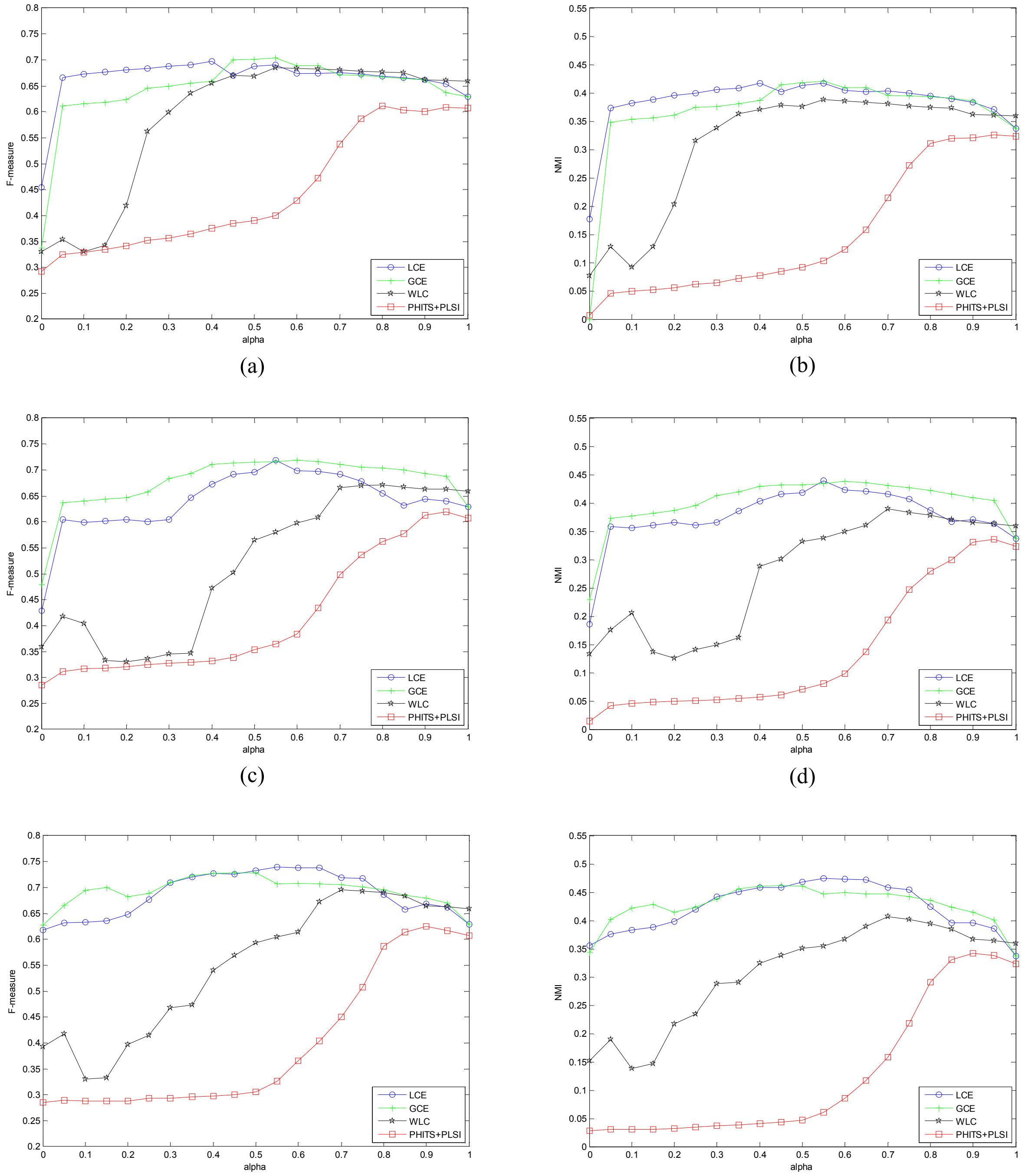

Figure 6 – Results on Citeseer dataset using inlinks (a-b), outlinks (c-d) and inlinks+outlinks (e-f)
(LCE: Local Content Enrichment; GCE: Global Content Enrichment; WLC: Weighted Linear Combination)