T. Dkaki[a,b], Q.D. Truong[a,c], J. Mothe[b], P-J. Charrel[a]
a GRIMM-ISYCOM Université Toulouse 2 France
b Institut de Recherche Informatique de Toulouse France
c CanTho University VietNam

A new method for information retrieval based on graph comparison

Pages: 3-7

Current Research in Information Sciences and Technologies

Multidisciplinary Approaches to Global Information Systems

Vicente P. Guerrero-Bote (Editor)

Proceedings of the I International Conference on Multidisciplinary
Information Sciences and Technologies, InSciT2006
Mérida - SPAIN
October, 25th-28th, 2006

ISBN-10 CD-ROM: 84-611-3106-1
ISBN-13 CD-ROM: 978-84-611-3106-8

ISBN-10 Whole Edition: 84-611-3103-7
ISBN-13 Whole Edition: 978-84-611-3103-7

ISBN-10 Volume I: 84-611-3104-5
ISBN-13 Volume I: 978-84-611-3104-4

Open Institute of Knowledge
(Instituto Abierto del Conocimiento)
Antonio Álvarez 6
06005 Badajoz, SPAIN
Phone: +34 924221935
Fax: +34 924221935
http://www.instac.es

A new method for information retrieval based on graph comparison

T. Dkaki[a,b], Q.D. Truong[a,c], J. Mothe[b], P-J. Charrel[a]
[a] *GRIMM-ISYCOM Université Toulouse 2 France*
[b] *Institut de Recherche Informatique de Toulouse France*
[c] *CanTho University VietNam*

We propose a new method for information retrieval (IR). The main goal of this method is to enhance the core IR-process of finding relevant and only relevant documents in a set of documents. More precisely the method aims at increasing precision at top ranked documents. It is founded on previous works on graph matching and acknowledges the significance of structural similarity in making analogies. The method we propose is based on graph vertices comparison and involves recursive computation of similarity. It extends the principle stating that the resemblance of two vertices can be computed over the similarities of the vertices to which they are connected and takes into account the concept of similarity propagation within a graph. First results from conducted experiments show the feasibility and the effectiveness of our approach. Indeed, our method highly outperforms the vector-based cosine model and more than doubles the precision until top sixty returned documents.

Keywords: Graph, Graphs comparison, information retrieval.

1 INTRODUCTION & RELATED WORKS

Searching for relevant information is a hard task and deciding whether or not a piece of information can fulfill someone's needs is somewhat complex. This highly demanding task mainly implies finding out how well a given document or chunk of text matches a user's query. Therefore, the main question is about assessing how theses two items –a document and a query- are similar.

This concept of similarity is difficult to circumvent in IR as it is generally the case in large areas related to cognition [1][2]. Indeed, recognition, clustering and categorization, case-based reasoning, generalization… are processes in which similarity intervenes as a constraint. Still, there is nothing such as a universal similarity assessment that can be measured straightforwardly [3]; it is always necessary to define in what regards two things/items are similar. Any similarity measurement is, therefore, model -conceptual space representation- dependent [4]. Many authors define similarity at two levels: the surface level and the structural level. Surface similarity is defined as an attribute-oriented function while structural similarity is defined as a relation-dependent function.

Our interest in structural similarity originates from the fact that nowadays IR systems generally handle huge amount of information and thus are more expected to perform well with respect to precision and that several cognitive psychology studies [5] suggest that structural similarity favours precision while surface similarity –as used in most IR models- favours recall over precision.

Features used in structural similarity computation are mainly relations. We are naturally interested in graph theory as graphs are common representations that can capture the structure and thus model a wide range of relational data and knowledge. Moreover, graph theory already plays a major role in many specific domains such as Web information retrieval [6][7]; text information retrieval [8][9], social networks analysis [10][11] and science citation and co-citation networks analysis [12]. Computing similarity based on graph structure has also been explored in the specific context of database schema matching [13]. [14] points out that the Web as a citation graph is structurally similar to the two-node graph *hub* → *authority* and expresses the approach of hub and authority analysis in [15] as a graph mapping issue. Basic approaches -using graphs- are based on similarity of immediate neighbors of a given node pair while more sophisticated approaches such as SimRank [12] are based on the whole graph structure.

In this paper, we further investigate these aspects and propose a precision driven method for information retrieval (IR). This method is based on graph vertices comparison. It is inspired by previous works dealing with both graph matching in discreet mathematics and similarity studies in cognitive psychology. The overall goal of this method is to enhance the core IR process of retrieving relevant information from a set of documents -where relevance is defined as an end-user's satisfaction measurement with respect to the needs expressed in queries.

We use bipartite graphs for document and query representation. Graph vertices are of two types: documents and queries, and indexation terms. Graph edges connect indexation terms to the documents and the queries they represent. The resulting IR graph-model facilitates the use of structural similarities in the process of matching documents and queries. This process becomes a graph vertices comparison issue. Thus, getting relevant documents becomes a search for document nodes similar to a query vertex.

Graph matching and comparison theories also give the theoretical bases for using latent similarities of both documents and terms. This is done by exploiting the fact that the similarity between two nodes from two graphs can be computed as a recursive and iterative function that enhances an initially given similarity measure. At each step, the computational method reevaluates the similarity measure between two nodes in light of contextual information by determining a similarity score between their sets of adjacent nodes.

The remainder of this paper is organized as follows. Section 2 presents our main approach in lights of those presented in [15] and [14], exposes primary tests and puts forward improvements to the first proposed methods. Section 3 discusses experimental results and section 4 suggests some perspectives.

2. APPROACH

The starting point of our approach is the method applied in Web information retrieval proposed in [15] and generalized in [14]. We improved these methods in order to properly take into consideration the case of self similarity as needed in Information Retrieval Systems (IRS). The resulting IR graph-model facilitates the use of structural similarities in the process of matching documents and queries which becomes a matter of graph vertices comparison.

In [15], Kleinberg proposed a method that allows the computation of hub and authority scores of hyperlinked pages as in (1).

$$h_{p_j} = \sum_{p_i:\ p_i\ references\, p_j} a_{p_i} \quad ; \quad a_{p_j} = \sum_{p_i:\ p_i\ references\, p_j} h_{p_i} \qquad \textbf{(1)}.$$

h_{pi} and a_{pi} are respectively the score of hub and the score of authority of page$_i$

Web search engines use these scores to increase their accuracy. In our approach, the computation of hub and authority scores is seen as a graph comparison problem [14] where the Web, as a citation graph, is compared to the two-nodes directed graph hub → authority. Indeed (1) can be expressed as follows:

$$\begin{bmatrix} h_{p_1} & a_{p_1} \\ \vdots & \vdots \\ h_{p_n} & a_{p_n} \end{bmatrix} = B \begin{bmatrix} h_{p_1} & a_{p_1} \\ \vdots & \vdots \\ h_{p_n} & a_{p_n} \end{bmatrix} \begin{bmatrix} 0\,1 \\ 0\,0 \end{bmatrix}^T + B^T \begin{bmatrix} h_{p_1} & a_{p_1} \\ \vdots & \vdots \\ h_{p_n} & a_{p_n} \end{bmatrix} \begin{bmatrix} 0\,1 \\ 0\,0 \end{bmatrix}$$

(2) B is the adjacency matrix of the web graph Matrix $\begin{bmatrix} 0\,1 \\ 0\,0 \end{bmatrix}$ is obviously the adjacency matrix of the two nodes graph hub→authority

More generally, when considering two structurally similar graphs, one to analyze -the target- and the other playing a role of model -the source-(fig 1), we can map the first onto the second, in a transfer-like approach [5], by identifying a context sensitive similarity measure between their sets of nodes. The similarity score s_{ij} between vertex *i* of target graph and vertex *j* of source graph can be computed as follows:

$$S_{ij} = \sum_{r:(r,i)\in E_B,\, t:(t,j)\in E_A} S_{rt} + \sum_{r:(i,r)\in E_B,\, t:(j,t)\in E_A} S_{rt} \qquad \textbf{(3)}$$

E_A, E_B are the arc sets of target and source graphs.

We can rewrite (3) in a matrix formulation manner as:

$$S_{k+1} = BS_k A^T + B^T S_k A \qquad \textbf{(4)}$$

B and A are the adjacency matrices of target graph and source graph

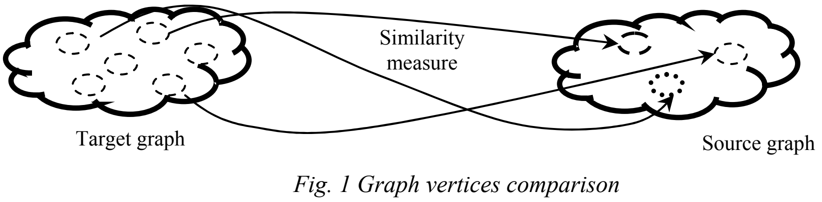


Fig. 1 Graph vertices comparison

(4) defines the similarity between nodes as a reflexive and recursive function. This triggers two fundamental questions related to the algorithm convergence and to the best initial similarity values (S_0).

<u>Convergence:</u> Convergence of (4) is uncertain. This problem on convergence can be overcome by normalizing the similarity matrix *S* at each iteration step. (4) is then rewritten as follows:

$$S_{k+1} = \frac{BS_k A^T + B^T S_k A}{\left\| BS_k A^T + B^T S_k A \right\|_F} \qquad \textbf{(5)}$$

$\| \|_F$ is the Frobenius norm

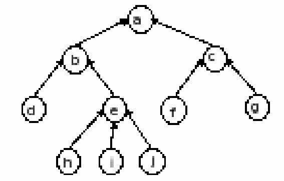
In this case, whatever the initial similarity values, the sequence admits two adherence values: one limit for series S_{2k} and another for series S_{2k+1} (see [14] for proof). The limit of sub-series S_{2k} is used as the similarity matrix between vertices of source and target graphs.

Initialization: There are two possible cases regarding how to choose an initial similarity matrix S_0. These cases are related to the a priori awareness of any resemblance between vertices of the two graphs. If there is no previously known information then it seems natural that all node pairs must be associated to the same score of similarity -e.g. 1- thus S_0 is a matrix full of ones. Otherwise, previously known similarity scores -e.g. some attributes dependent surface similarity- can be used to build matrix S_0.

2.1. Preliminary tests & Method enhancement

Initial tests give broadly good results. The method shows quite satisfactory mapping of target graph onto source graph. However, in specific cases, the results we obtain, although they are fully explicable (see Fig. 2), seem somewhat unnatural. This calls for an improvement of the method.

In the following example, the graph represented in fig. 2 is compared to the graph *hub→authority*. The initial similarity matrix is the matrix full of ones. The obtained results are unsatisfactory if not to say odd. They oppose commonsense -or at least the results of an in-degrees/out-degrees analysis. Indeed, examining the table in fig. 2, we notice that vertices a, d, f, g, h, i and j get the same authority score.



	a	b	c	d	e	f	g	h	i	j
Hub	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.2
Authority	0.0	0.0	0.0	0.0	0.86	0.0	0.0	0.0	0.0	0.0

Fig. 2 a graph used for preliminary test purpose (left) and results table from comparison with graph hub→authority

There are reasons that can explain these results. Perhaps the most important one is that our method does not take into account the notion of similarity inheritance –or flooding similarity. When considering a feature such as authority, the inheritance must somehow play a role in similarity evaluation over a graph set of nodes.

(5) needs to be modified to comply with the similarity inheritance principle. The propagation and the retro-propagation of similarities via a graph are considered as a "flooding" similarity [13] which can be expressed as follows:

$$S_{k+1}(i,j) = S_k(i,j) + \sum_{a:(a,i)\in A, b:(b,j)\in B} S_k(a,b)w((a,b),(i,j)) + \sum_{a:(i,a)\in A, b:(j,b)\in B} S_k(a,b)w((a,b),(i,j)) \quad (6)$$

W is a weighting function.

In our approach, the propagation of similarity can be considered a transitive graph closure of target and or source graph in a more sophisticated approach than in [13]; we add a more wide-ranging function of gradual attenuation of "inheritance over generations". Note that in [13] this attenuation is solely depth sensitive. In our approach, adjacency matrices A and B of source and target graphs are modified to meet the goal of flooding similarity attenuation as follows:

$$A \leftarrow A + \sum_{n=2}^{\infty} f_A(n)\frac{A^n}{\|A^n\|} \quad ; \quad B \leftarrow B + \sum_{n=2}^{\infty} f_B(n)\frac{B^n}{\|B^n\|} \quad (7)$$

$f_A(n) = \alpha^n$ and $f_B(n) = \beta^n$ where α and β are positive constants lower than 1 appear to be a natural instantiating attenuation functions. The influence of 'generations' exponentially decreases with path depth. By using (7), we also consider that inheritance from a "forebear" node of n^{th} generation of a vertex v is proportional to the number of paths of length n separating this forebear from node v. This could be somewhat questionable. We, then, refine this concept of inheritance to break this proportionality relationship. Adjacency matrices of target and source graphs are reformulated as follows:

$$A \leftarrow A + \sum_{n=2}^{\infty} f_A(n)g_A\left(\frac{A^n}{\|A^n\|}\right) \quad ; \quad B \leftarrow B + \sum_{n=2}^{\infty} f_B(n)g_B\left(\frac{B^n}{\|B^n\|}\right) \quad (8)$$

gA and gB are monotonically increasing functions between [0, 1] and [0,1]. They can be exponential or a staircase functions.

The use of these changes gives quite satisfactory results. The disadvantages depicted above are overcome as shown in the following table.

	a	b	c	d	e	F	g	h	i	j
Hub	0.0	0.12	0.12	0.15	0.15	0.05	0.05	0.27	0.27	0.27
Authority	0.36	0.37	0.09	0.0	0.65	0.0	0.0	0.0	0.0	0.0

Fig. 3 Enhanced results from comparison of graph in figure 2 with graph hub→authority

2.2. Graph Self comparison & Information Retrieval

The core IR process is the retrieval of relevant information in a set of documents. Relevance is defined as a measurement of documents' concordance with the user's needs expressed in a query. Using graph comparison for IR assumes that we look for document vertices similar to a given query node. This is a search for similar nodes of the same graph which implies graph self-comparison where target and source graphs are the same. Unfortunately, when assessing graph self-comparison, there are cases where *s(i, j) ≥ s(i, i)* –see figure 4. In fact, this opposes a necessary condition that a similarity measure must fulfill. The measure we obtain is positive defined -∀(i, j), s(i, j) ≥ 0-, symmetric -∀(i, j), s(i, j) = s(j, i)- but it does not always verify ∀(i, j), s(i, i) = s(j, j) ≥ s(i, j).

Satisfying this condition is, of course, not mandatory as very common "similarity" measures in IR such as the dot product do not fulfill this prerequisite. Nevertheless, to avoid this "disadvantage", we normaliz the similarity matrix (S_{AB}) by dividing each value $S_{AB}(i, j)$ by the product of self-similarity $S_{AA}(i, i)$ of vertex i in graph A and $S_{BB}(j, j)$ of vertex j in graph B. The final algorithm is described in paragraph 2.3.

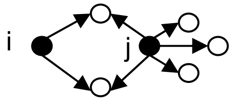


Fig. 4 graph self comparison: when analysing the graph above one can deduce that s(i, j) ≥ s(i, i)

2.3. Algorithm

```
$S_0 \leftarrow 1$
$k \leftarrow 0$
$A \leftarrow A + \sum_{n=2}^{\infty} f_2(n) g_2(\frac{A^n}{\|A^n\|})$
$B \leftarrow B + \sum_{n=2}^{\infty} f_1(n) g_1(\frac{B^n}{\|B^n\|})$
Re peat
    $S_{AA_{k+1}} \leftarrow \frac{AS_{AA_k}A^T + A^T S_{AA_k} A}{\|AS_{AA_k}A^T + A^T S_{AA_k} A\|_F}$
    $S_{BB_{k+1}} \leftarrow \frac{BS_{BB_k}B^T + B^T S_{BB_k} B}{\|BS_{BB_k}B^T + B^T S_{BB_k} B\|_F}$
    $S_{AB_{k+1}} \leftarrow \frac{BS_{AB_k}A^T + B^T S_{AB_k} A}{\|BS_{AB_k}A^T + B^T S_{AB_k} A\|_F}$
    $k \leftarrow k + 1$
Until convergence is achieved for k even
$S_{AB} \leftarrow \bullet \frac{S_{AB} \bullet * S_{AB}}{diag(S_{AA}) \bullet * diag(S_{BB})^T}$
Output  $S_k$  ($k$ is even) as similarity matrix
```

*Fig. 5 Algorithm for similarity matrix computation;. ● * and ● — are term to term matrix multiplication and division*

3. Experiments & results

For convenience reasons, we use data from TREC 2004 Novelty Track [16][17] to conduct our experiments. Information granularity is defined at the sentence level. Sentences are extracted from relevant documents and the task consists of retrieving those sentences that are relevant. In TREC 2004, 50 topics were used from the TREC collection. For each, the NIST selected relevant documents with a maximum of 25 documents per topic, which were given to participants with sentences marked-up. After runs were submitted, NIST evaluators decided which sentences were relevant.

For each topic, we construct a bipartite graph in which nodes are sentences and terms. Edges go from sentence node to term node and are weighted following tf*idf approach. Sentence retrieval is performed by using our method -algorithm described in section 2.3- and a cosine-based space vector model.

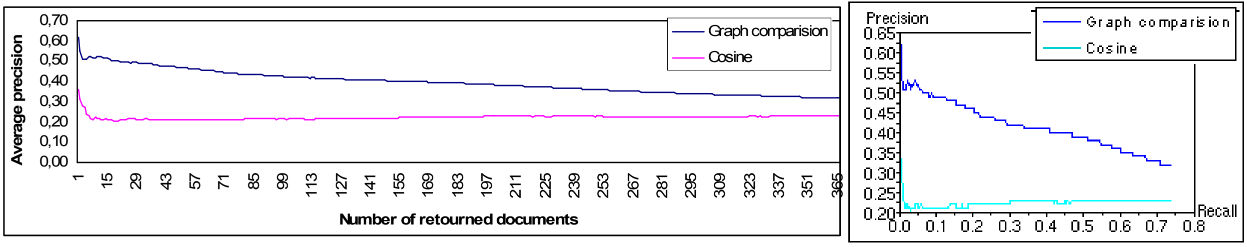


Fig. 6 results comparison over the Trec'2004 50 topics. This figure shows the average precision at top N returned documents (left) and the reall-precision chart (right) for both our method and the cosine model.

The experiments we conducted show the practicability of our approach. Moreover, the results reported in fig. 6 show the full potential of our method. Indeed, our method highly outperforms the vector-based cosine model. In fact, it more than doubles the precision until top 60 returned documents.

4. CONCLUSION

In this paper, we proposed a new model for computing similarity scores between vertices of two graphs by enhancing methods previously suggested in [12], [13], [14] and [15]. The proposed method has been tailored to suit the case where the two compared graphs are identical. This paved the way to our proposal of a new IR model. This model sees documents and queries along with indexation terms as vertices of a bipartite graph. The IR comparison task is achieved as a graph comparison process. Returned documents are the graph nodes satisfactorily similar to the query node.

The experiments undertaken showed that our method highly outperformed the vector-based cosine model. Also, there are hints that our method will offer large perspectives that go beyond traditional text retrieval as argued in this paper. Involved domains are XML retrieval –which can be achieved by using multipartite labeled graphs-, data mining and social networks analysis especially within the framework of Interactionism [18], where prominent room in explaining social facts is given to the relations between individuals. In the last domain, our method promises to go beyond the traditional analysis method by allowing network-analysis customization and by offering alternatives to the beaten path in traditional network analysis. It consists in defining an easily interpretable network –such as the 'buy from' network *Client→Retailer→Wholesaler*- which will play the source graph role and on comparing it to a source network to be analyzed -such as a large instantiated 'buy from' network. Another major perspective consists in extending the set of possible target graphs by considering semantically richer graphs such as labeled graphs or hyper-graphs. This will make possible to better take into account the complexity of links within social networks.

REFERENCES

[1] D. Gentner, M. J. Ratterman, K. D. Forbus, 'The roles of similarity in transfer: Separating retrievability From inferential soundness', Cognitive Psychology, 25, 524-575, 1993.

[2] D.L. Medin, R.L. Goldstone, and D. Gentner, 'Similarity involving attributes and relations: Judgments of similarity and difference are not inverses', Psychological Science, 1(1): 64-69, 1990.

[3] N.Goodman, 'Seven strictures on similarity. In Goodman, N. (ed.) Problems and Projects, pp. 437-447. Indianapolis and New York: Bobbs-Merrill, 1972.

[4] P. Gärdenfors, 'Conceptual Spaces. The Geometry of Thought', Cambridge, Mass.: MIT Press, 2004.

[5] D. Bracke, 'Vers un modèle théorique du transfert: les contraintes à respecter', Revue des sciences de l'éducation, XXIV(2) :235-266, 1998.

[6] M. Henzinger, 'Link Analysis in Web Information Retrieval', Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2000.

[7] M. Sahami, V. Mittal, S. Baluja, H. Rowley, 'The Happy Searcher: Challenge in Web Information Retrieval', In Trends in Artificial Intelligence, 8th Pacific Rim International Conference on Artificial Intelligent (PRICAI), 2004.

[8] M. M. Gómez, A. Gelbukh, A. L. López, 'Comparison of Conceptual Graphs', Proc. 1st Mexican International Conference on Artificial Intelligence, 2000. Lecture Notes in AI N 1793, ISSN 0302-9743, Springer, pp. 548-556.

[9] Y. Quintana, M. Kamel, A. Lo, 'Graph-based retrieval of information in hypertext systems', Proc. of the 10th annual international conference on Systems documentation, pp. 157-168, 1992.

[10] L. C. Freeman, 'Centrality in Networks: Conceptual Clarification', Social NetWork 1, pp. 215-239, 1979.

[11] M. E. J. Newman, 'Random graphs as models of networks', in Handbook of Graphs and Networks, S. Bornholdt and H. G. Schuster (eds.), Wiley-VCH, Berlin 2003.

[12] G. Jeh, J. Widom, 'SimRank: a measure of structural-context similarity', Proc. of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 538-543, 2002.

[13] S. Melnik, H. Garcia-Molina, E. Rahm, 'Similarity Flooding : A Versatile Graph Matching Algorithm and its Application to Scheme Matching', Proc. of the 18th ICDE Conference, 2002.

[14] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren,'A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching', SIAM Rev. 46(4):647-666, 2004.

[15] J. M. Kleinberg, 'Authoritative Sources in a Hyperlinked Environment', Journal of the ACM, 46(5):604-632, 1999.

[16] TREC, http://trec.nist.gov

[17] T. Dkaki, J. Mothe, 'Trec Novelty Track At IRIT-SIG', Text Retrieval Conference, 2004.

[18] Blumer, Herbert, 'Symbolic Interactionism: Perspective and Method', Englewood Cliffs, NJ/ Prentice-Hall, 1969.