Asociación Argentina de Mecánica Computacional

Mecánica Computacional Vol XXXI, págs. 3411-3422 (artículo completo)
Alberto Cardona, Paul H. Kohan, Ricardo D. Quinteros, Mario A. Storti (Eds.)
Salta, Argentina, 13-16 Noviembre 2012

STRONGLY CONNECTED COMPONENTS DETECTION IN OPEN DIRECTORY PROJECT GRAPHS

Eduardo Xamena[a,b,c,e], **Nélida B. Brignole**[a,b] **and Ana G. Maguitman**[c,d]

[a]*Planta Piloto de Ingeniería Química (UNS-CONICET) - Camino la Carrindanga km 7, Bahía Blanca, Argentina.*

[b]*LIDeCC - Laboratorio de Investigación y Desarrollo en Comp. Científica.*

[c]*Knowledge Management and Information Retrieval Research Group.*

[d]*LIDIA - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial.*

Departamento de Ciencias e Ing. de la Computación - Universidad Nacional del Sur, Av. Alem 1253, Bahía Blanca, Argentina. {ex,nbb,agm}@cs.uns.edu.ar

[e]*Facultad de Ciencias Exactas - UNSa - Universidad Nacional de Salta, Av. Bolivia 5150, Salta, Argentina.*

Keywords: Strongly Connected Components Detection, ODP, Graphs.

Abstract. The Strongly Connected Components Detection Algorithm is a powerful tool for identifying representative subsets on big communities. By means of a graph representation, smaller groups of nodes related to each other by some particular feature can be obtained. In this work, we propose to apply this algorithm to the Open Directory Project (ODP) graph to identify topical communities of web pages. The challenge of this research work is to find groups of topics strongly connected to each other, which will enable us to perform further similarity studies between web pages from ODP. We discuss some findings observed from the application of this algorithm to the ODP graph.

1 INTRODUCTION

Nowadays the need for information within large document repositories is very common. One of the most frequent cases is the search for specific content on the Internet. For this particular case, there are various search platforms that give us more or less accurate results for each information need, such as Google[1] or Bing[2]. There are also catalogs of websites, developed jointly by large groups of people around the world. Such is the case of the Open Directory Project (ODP)[3], which consists of a large directory of websites organized by topics, with a hierarchical structure. Each page that is added to the directory should be added to a topic of the structure, and each topic can be subdivided into other topics. The aim of this work is the search and analysis of topical communities with a high degree of cohesion. These communities may be used in other studies as a basis for semantic similarity research in large document corpora.

The way to find these communities is the implementation of a directed graphs algorithm on the ODP structure, which may be useful for the task of recognizing them. This algorithm is the *Strongly Connected Components Detection* (SCCD) in directed graphs (Tarjan (1972)). Its application in the large graphs determined by the hierarchical and nonhierarchical relationships of ODP topics, may determine well-defined groups of topics that present common characteristics and serve as a basis for further experiments.

Representing document corpora by means of graphs can be very useful for different purposes. In Agirre et al. (2009) for example, a calculation of semantic similarity by using this concept is shown. Also in Boldi and Vigna (2004), a work on compression and rendering of large volumes of information on the structure of the web is carried out, and in Broder et al. (2000) various analyzes on Internet organization and its macro-structure of connectivity are performed.

Identifying thematic communities in large volumes of information can be a starting point for further research works. Recommendation systems, for example, can benefit from extensive use of these communities formed in the corpora (Belkin (2000); Balabanović and Shoham (1997); Akavipat et al. (2006)). There are also several areas of application that can benefit from the creation of these communities, such as web crawlers (Kumar and Vig (2009); Aggarwal et al. (2001); Chakrabarti et al. (1999)) and social networks (Schifanella et al. (2010); Lin et al. (2009)).

The following section details the basic aspects of graph theory for introducing the algorithm used in this work. Then the ODP project is described. After that, the process of implementing the algorithm on the directed graph structure of the ODP is explained. Finally, we express the respective conclusions and possible lines of inquiry to follow.

2 GRAPH THEORY

In this section we briefly explain some useful issues of graph theory and introduce the SCCD algorithm. These concepts are the theoretical basis of this work. The communities search on ODP that we carry out relies on these features.

2.1 Some basic concepts about graphs

A directed graph consists of a set of nodes, denoted V, and a set of edges, denoted E. Each edge is an ordered pair of nodes (u, v) representing a directed connection from u to v. A path from node u to node v is a sequence of edges $(u, u1), (u1, u2)...(uk, v)$. If this set of edges

[1] http://www.google.com/
[2] http://www.bing.com/
[3] http://dmoz.org/

exists we say that we can "walk" through the graph from u to v. Note that a path from u to v does not imply a path from v to u on a directed graph.

A *Strongly Connected Component* (SCC) on a directed graph is a set of nodes such that for any pair of nodes u and v in the set there is a path from u to v. A directed graph may have one or more SCC's. These components consist of disjoint sets of nodes of the graph. On figure 1 we can see a graph and its corresponding SCC's. These concepts are further analyzed in Broder et al. (2000).

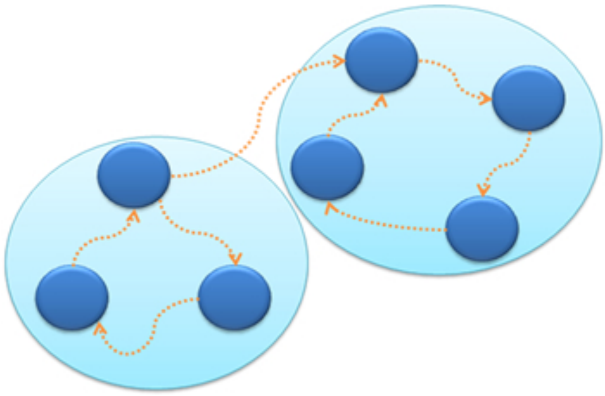

Figure 1: Strongly connected components of a graph.

2.2 Strongly Connected Components Detection Algorithm

In order to identify the SCC's of a directed graph, in Tarjan (1972) a very efficient algorithm is proposed. It is based on another algorithm that conducts a tour through the whole nodes of the graph. The latter algorithm is known as *Depth-First Search* (DFS), and consists on visiting systematically every node. Its denomination corresponds to the fact that this algorithm starts from one random node, selects an adjacent node, and looks for an adjacent to the last one, doing this until there are no more adjacent nodes, or all of them were visited previously. Thus, the mentioned algorithm prioritizes depth-first traversal rather than breadth-first traversal. Into the implementation, DFS uses a *Depth-First index* (DFI) number for labeling nodes in the order they are discovered. The pseudocode of this algorithm is:

```
Procedure DFST:
  Input: N, E
  Output: E_f (Final Edges of the DFS)

  i ← 1
  E_f ← ∅
  For every v ∈ N do: DFI(v) ← 0
  While exists some u ∈ N such that DFI(u) = 0 do
    DFS(i, u, N, E, E_f)
Procedure DFS:
  DFI(v) ← i
  i ← i + 1
  For every node v' adjacent to v do:
    If DFI(v') = 0
    then
```

```
      E_f ← E_f ∪ {(v, v')}
      DFS(i, v', N, E, E_f)
    endif
  endfor
```

The first procedure covers all the nodes in the graph, while the second acts while there is a connection between a visited node and a new node. The result of this algorithm is represented by the set E_f that represents the ordered edges that should be traversed for performing a DFS on the graph.

The main concept of these procedures is used by the SCCD algorithm, since it traverses the nodes in a similar way. On its implementation, SCCD proceeds labeling the nodes with two weight numbers: DFI(v) and Q(v). The former is the one used in the DFS algorithm, and the latter corresponds to a value of the algorithm that helps on identifying the roots of the SCC's. For further detail on this question, see Tarjan (1972). Next, we show the pseudocode of the SCCD algorithm:

```
Procedure DFSCCG:
  Input:  i, v, DFI, N, E, P
  Input-Output:  j, CC

  DFI(v) ← i
  Q(v) ← DFI(v)
  i ← i + 1
  Put v in P
  Stacked(v) ← true
  For every node v' adjacent to v do:
    If DFI(v') = 0
    then
      DFSCCG(i, v', DFI, N, E, P, j, CC)
      Q(v) ← min(Q(v), Q(v'))
    else
      If DFI(v') < DFI(v)andStacked(v')
      then
        Q(v) ← min(Q(v), DFI(v'))
      endif
    endif
  endfor
  If Q(v) = DFI(v)
  then
    Unstack all the elements from P until reaching v
    Store all the unstacked elements in CC(j), including v
    Unstack v
    Set Stacked(u) ← false, for everynode in CC(j)
    j ← j + 1
    {elements strored in CC(j) make up the j − thSCC
    of the graph with root v}
  endif
```

```
Procedure TCCD:
  Input:  N,E
  Input-Output:  CC

  i ← 1
  j ← 1
  P ← ∅
  CC ← ∅

  For every v ∈ N do:
    DFI(v) ← 0
    Stacked(v) ← false
  endfor
  While exists some u such that DFI(u) = 0 do:
    DFSCCG(i, u, DFI, N, E, P, j, CC)
  endwhile
```

Both algorithms implement a modified DFS that identifies SCC's in a graph. The first traverses all the connected nodes from a given node, and obtains the corresponding SCC's. The second algorithm looks for every node that remains unvisited after running the first one. The result of the SCCD algorithm is the set CC which contains the node sets for every SCC found.

3 THE OPEN DIRECTORY PROJECT

There are countless websites that are visited daily by millions of people for various purposes. The information needs of these users are varied, and seeking information on a topic would be a very difficult task if they were not any tools such as Internet search engines or web directories, that help users on those issues. One of these directories is the ODP project (figure 2), which is widely exploited by users and research projects. Some of the research projects mentioned use ODP for training and testing automatic classifiers (Biro et al. (2008); Gauch et al. (2009)), as the starting point to collect thematic material by topical crawlers (Chakrabarti et al. (1999); Menczer et al. (2004)), as a framework to understand the structure of content-based communities on the Web (Chakrabarti et al. (2002)), to implement Information Retrieval evaluation platforms (Beitzel et al. (2003); Maguitman et al. (2010)), to understand the evolution of communities in P2P search (Akavipat et al. (2006)), to define hierarchically-informed keyword weight propagation schemes (Kim and Candan (2007)) and to evaluate the emergent semantics of social tagging (Markines et al. (2009)), among other applications.

3.1 Organizing and Representing the Directory

ODP is an open project maintained by thousands of users around the world, that catalog pages in different areas of interest. Each one of these areas identifies a topic, and these topics are organized into a graph structure. The nodes of this graph are the different topics of ODP, and the edges are relations between these topics, i.e. the pages contained therein. These relationships can be of three types:

- **Hierarchical**: These correspond to the main directory hierarchy, through which the site categories are organized.

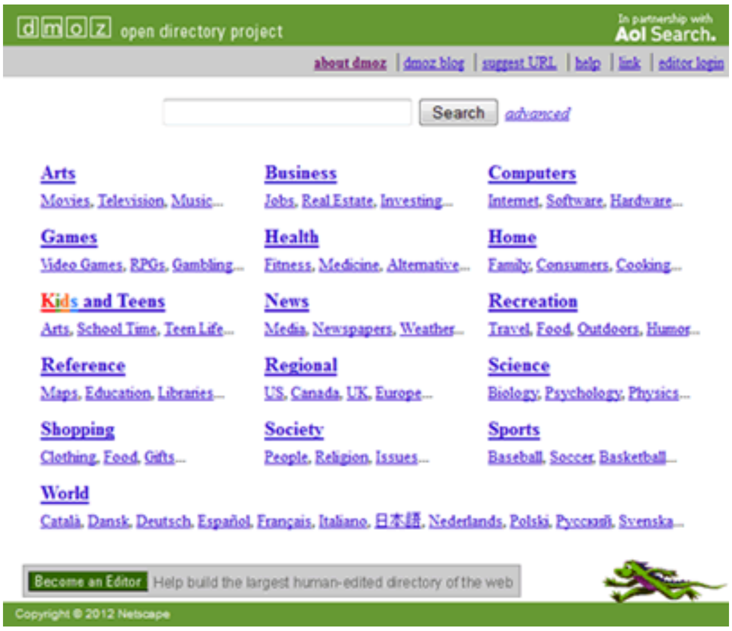

Figure 2: Screenshot of the DMOZ ODP project site.

- **Symbolic**: Edges of the same level as hierarchical, intended to meet the taxonomic relationships of some alternative organization of the site. This happens because many times we can characterize more than one topic as the immediate ancestor of the same subtopic, which would be impossible to achieve in a purely hierarchical structure.
- **"See also"**: Edges that connect related topics. They have a lower significance level than the edges of the other two classes.

An example of a portion of the ODP graph is shown in figure 3.

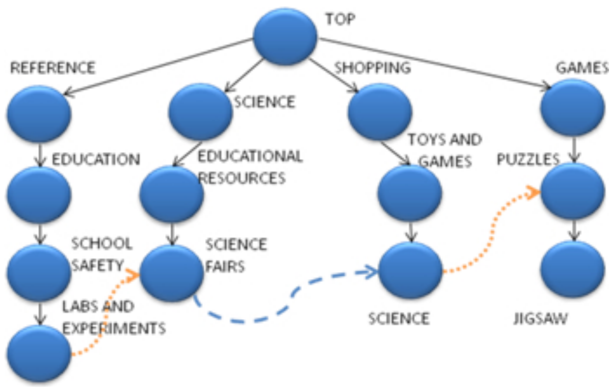

Figure 3: A portion of the ODP graph, showing the different kinds of links between topics.

In this graph, hierarchical links are represented by a black full line, the symbolic links with blue stripes line, and "see also" links with an orange dotted line.

Because of the amount of web pages that exist in the directory, and as a consequence of this the amount of present topics therein, the graph and its representation corresponding matrix are extremely large. To represent a version of ODP, we used matrices that were close to 600.000 rows and 600.000 columns in size.

4 SCCD ALGORITHM IMPLEMENTATION ON ODP

The implementation of SCCD algorithm on the ODP graph was carried out by using a library developed for other purposes. Such library is a part of the *Direct Method for Observability Analysis of Chemical Plants Instrumentation Design* (Ponzoni (2001)). On its application area, the SCCD algorithm performs a rearrangement on a block of the adjacency matrix of a large sparse system of equations. Once it is applied, the system is subdivided in assignment blocks for obtaining the values of some variables. The library that implements the algorithm is written in the *C programming language*, and is a separate package of the Direct Method mentioned here.

4.1 Data Structures

In order to make possible the implementation of the SCCD algorithm on the ODP graph, it was necessary to translate the associated data. This was done with a version of the ODP made up of 600.000 topics approximately, the same one that was used in Xamena et al. (2011). The data input format accepted by the library is a sparse representation of the digraph's adjacency matrix and consists of an enumeration of the nonzero elements of the matrix that represents the digraph. In the ODP case, these nonzero elements match with the relations between topics. Each kind of relation mentioned in the corresponding section is codified in a matrix that represents the edges of the ODP graph. The columns and rows of that matrix correspond to the topics on ODP, and every relation from a topic identified with a row to another topic identified with a column is represented as a nonzero value in the respective matrix cell.

4.2 Strongly Connected Components Found

We ran the SCCD algorithm implementation mentioned above on the ODP associated data structure. Table 1 summarizes the results obtained. There we can see the number of identified SCC's grouped by quantity of nodes that make up each component. For example, 12 SCC's of size 22 were found, and 8 SCC's of size 25, while no SCC of size 35 was found.

The information of table 1 shows that there exists a huge strong component of size 279519. It is quite likely to find topics within this component that should not be related. This key aspect could mean that we have to be careful when taking into account topic relations determined by this component, given that the nodes that make it up represent almost a half of the whole nodes in the ODP graph.

Another interesting aspect is the existence of too many isolated nodes, each one forming a component of one node. That could give rise to further studies on the nodes involved, in order to determine if there is really not any possible relation to take into account among one of these nodes and some others. The only reason because these topics are isolated is that the algorithm did not find a cycle involving them. Otherwise, they should be a part of some other component.

As in the case of the components of one node, where just adding an edge linking a node to another one we could include an isolated node into some SCC, if we remove a key edge involved in the formation of a SCC it could result in the elimination of many nodes that are included within that SCC. By observing that, we could suppose that some edges involved in the big strong component of 279519 nodes identified in table 1 should be removed, given that they do not add significant information about the relation between topics. If we look at the edge types that are present in ODP, it is remarkable that the "see also" edges could be less significant than the hierarchical and symbolic edges in terms of relation between topics, owing to its origin. Therefore, if we withdraw the edges of this type from the ODP representation for

Table 1: Number of Strongly Connected Components found, ordered by size.

Size	**Number**	**Size**	**Number**	**Size**	**Number**
1	260702	21	8	46	1
2	5305	22	12	48	1
3	1524	23	5	49	1
4	632	24	7	51	1
5	381	25	8	54	1
6	225	26	1	60	1
7	167	27	3	64	1
8	98	28	3	70	1
9	70	29	4	78	1
10	77	30	4	81	1
11	43	31	4	83	2
12	48	32	2	85	1
13	28	33	2	86	1
14	31	34	1	89	2
15	25	37	1	98	1
16	18	41	2	101	1
17	12	42	1	123	1
18	12	43	1	279519	1
19	6	44	2		
20	8	45	1		

Table 2: Id numbers and Roots of topics for an example SCC with a new significant relation found.

Id	Complete Route
502266	Top/Computers/Software/Typesetting/TeX/Plain_TeX/Macros
155677	Top/Science/Math/Publications/Style_Files
300975	Top/Computers/Software/Typesetting/TeX/Macros
4795	Top/Computers/Software/Typesetting/TeX/LaTeX/Macros

Table 3: Id numbers and Roots of topics for an example SCC with a new significant relation found.

Id	Complete Route
91990	Top/Regional/North_America/United_States/Louisiana/ Travel_and_Tourism/Travel_Services/Tour_Operators
380010	Top/Recreation/Outdoors/Guides_and_Outfitters/North_America/ United_States/Louisiana
221405	Top/Recreation/Outdoors/Hunting/Guides_and_Outfitters/North_America/ United_States/Louisiana
34931	Top/Regional/North_America/United_States/Louisiana/ Recreation_and_Sports/Fishing_and_Hunting/Guides_and_Charters

obtaining SCC's, maybe we could get smaller components derived from the large component identified in this work, and with more significant relations between topics.

Going deeper into the analysis of the obtained SCC's, we can see examples of relations that could add significance to our representation, e.g. for the implementation of a semantic similarity measure. A case of such a relation is the one returned by the algorithm in a SCC which involves the topics in table 2. As a consequence of obtaining that SCC, we could say that all of these topics are related one with each other. Then, we have the existence of a relation between topics 502266 (Top/Computers/Software/Typesetting/TeX/Plain_TeX/Macros) and 155677 (Top/Science/Math/Publications/Style_Files), that does not exist in the ODP graph, but could add significance to its model.

Although we can find new relations that keep the coherence of the model, some derived associations between topics are not meaningful. An example of this situation is the relation derived by the SCC depicted in 3 for topics 91990 (.../Travel_and_Tourism/Travel_Services/Tour_Operators) and 34931 (.../Recreation_and_Sports/Fishing_and_Hunting/Guides_and_Charters). Since this relation could be questionable, the entire SCC is not so reliable.

There is a plethora of examples like the two described above, and it suggests that some additional work must be done with the obtained information, in order to use it for other research purposes. Another interesting issue that requires further analysis is the frequency distribution of the SCC's sizes. Figure 4 sheds some light into this issue. This chart shows the number of components obtained according to the number of nodes within each one, in logarithmic scale. The horizontal axis represents the sizes, and the vertical axis the number of components found. At first sight, there seems to be a relation between the number of components and their size that could be useful to induce many conclusions. But a deep statistical analysis is required for these purposes.

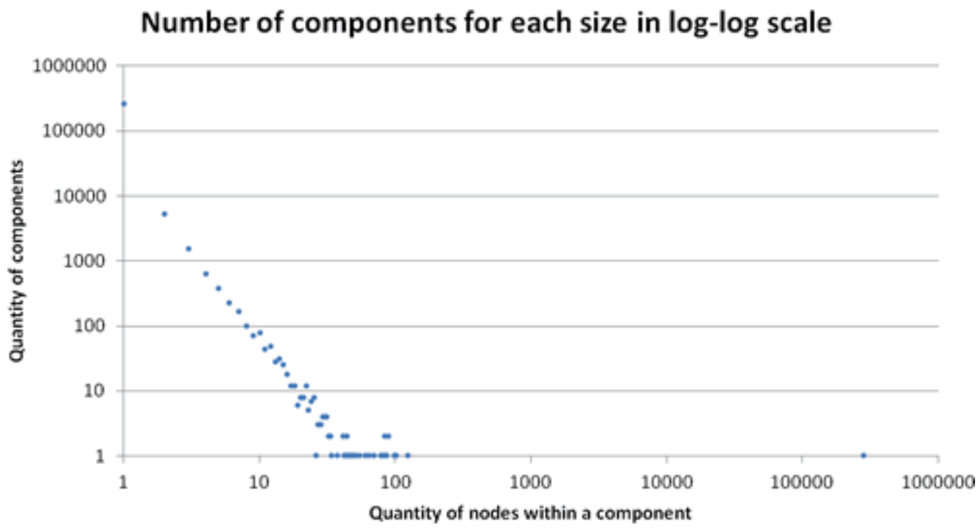

Figure 4: Chart for the number of SCC's found per size in number of topics.

5 CONCLUSIONS

We have implemented the Strongly Connected Components Detection Algorithm in the ODP graph structure. The main objective of our work was to identify topical communities of interest for further research tasks and experiments. The topical communities were efficiently obtained by using a C library adjusted for this purpose. It is remarkable that the program arrived to the final result, including the automatic achievement of a summary of the components found, in less than one hour. It dealt with a data structure of 2.269.866 edges and obtained 269.504 strong components. After that, it elaborated the mentioned summary that reflects the account of components grouped by their size.

The algorithm implementation resulted in the finding of a great number of components of the ODP graph that were summarized in the corresponding section. Some highlights on two representative topic groups are shown. For the first case study, the existence of a new relevant topic relation is discovered, due to the strong component identification. For the second, it is explained that some relations of the new topical groups may not be very relevant; hence it is mandatory to take additional measures on those relations, in order to include only the significant information of these groups. In other words, as expressed in Xamena et al. (2011), it is not enough to work on structural aspects of this kind of directories only, but it would be better to add information of the content and links between the pages contained within each topic to enhance and weight the identified new relations.

With the attained topical communities, there are several research lines to follow. Firstly, the SCCD algorithm could be applied to other data structures representing different information corpora, e.g. WordNet[4] or GeneOntology[5]. Secondly, some strong components found in this article could be studied deeply, in order to determine the degree of relation among the involved topics. Besides, we could get improved results by working with the data structure of this paper, but not including the "see also" edges of the ODP directory. An additional interesting work could be a statistical analysis of the components, following the relation between their sizes and the quantity of components found for each. Maybe we could find a model that represents the behavior of the SCC's when the directory grows in size. Further experiments should be carried

[4]http://wordnet.princeton.edu/
[5]http://www.geneontology.org/

out for validating the information achieved in this work. This will be part of our future research.

REFERENCES

Aggarwal C.C., Al-Garawi F., and Yu P.S. Intelligent crawling on the World Wide Web with arbitrary predicates. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 96–105. ACM, New York, NY, USA, 2001. ISBN 1-58113-348-0. doi:10.1145/371920.371955.

Agirre E., Alfonseca E., Hall K., Kravalova J., Paşca M., and Soroa A. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27. Association for Computational Linguistics, Stroudsburg, PA, USA, 2009. ISBN 978-1-932432-41-1.

Akavipat R., Wu L.S., Menczer F., and Maguitman A.G. Emerging semantic communities in peer web search. In *Proceedings of the international workshop on Information retrieval in peer-to-peer networks*, P2PIR '06, pages 1–8. ACM, New York, NY, USA, 2006. ISBN 1-59593-527-4. doi:10.1145/1183579.1183581.

Balabanović M. and Shoham Y. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997. ISSN 0001-0782. doi:10.1145/245108.245124.

Beitzel S.M., Jensen E.C., Chowdhury A., and Grossman D. Using titles and category names from editor-driven taxonomies for automatic evaluation. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 17–23. ACM, New York, NY, USA, 2003. ISBN 1-58113-723-0. doi:10.1145/956863.956868.

Belkin N.J. Helping people find what they don't know. *Communications of the ACM*, 43(8):58–61, 2000. ISSN 0001-0782. doi:10.1145/345124.345143.

Biro I., Benczur A., Szabo J., and Maguitman A. A comparative analysis of latent variable models for web page classification. In *Proceedings of the 2008 Latin American Web Conference*, pages 23–28. IEEE Computer Society, Washington, DC, USA, 2008. ISBN 978-0-7695-3397-1. doi:10.1109/LA-WEB.2008.14.

Boldi P. and Vigna S. The WebGraph framework I: Compression techniques. In S.I. Feldman, M. Uretsky, M. Najork, and C.E. Wills, editors, *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 595–602. ACM, New York, NY, USA, 2004. ISBN 1-58113-844-X. doi:10.1145/988672.988752.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., and Wiener J. Graph structure in the web. *Computer Networks*, 33(1âĂŞ6):309 – 320, 2000. ISSN 1389-1286. doi:10.1016/S1389-1286(00)00083-9.

Chakrabarti S., Joshi M.M., Punera K., and Pennock D.M. The structure of broad topics on the Web. In *Proceedings of the 11th international conference on World Wide Web*, WWW '02, pages 251–262. ACM, New York, NY, USA, 2002. ISBN 1-58113-449-5. doi:10.1145/511446.511480.

Chakrabarti S., van den Berg M., and Dom B.E. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31:1623–1640, 1999.

Gauch S., Chandramouli A., and Ranganathan S. Training a hierarchical classifier using inter document relationships. *J. Am. Soc. Inf. Sci. Technol.*, 60:47–58, 2009. ISSN 1532-2882. doi:10.1002/asi.v60:1.

Kim J.W. and Candan K.S. Leveraging structural knowledge for hierarchically-informed keyword weight propagation in the web. In *Proceedings of the 8th Knowledge discovery on the web international conference on Advances in web mining and web usage analysis*, WebKDD'06, pages 72–91. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 3-540-77484-X, 978-3-540-77484-6.

Kumar M. and Vig R. Design of core: context ontology rule enhanced focused web crawler. In *Proceedings of the International Conference on Advances in Computing, Communication and Control*, ICAC3 '09, pages 494–497. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-351-8. doi:10.1145/1523103.1523201.

Lin H., Davis J., and Zhou Y. An integrated approach to extracting ontological structures from folksonomies. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 654–668. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02120-6.

Maguitman A.G., Cecchini R.L., Lorenzetti C.M., and Menczer F. Using Topic Ontologies and Semantic Similarity Data to Evaluate Topical Search. In C. von Lücken, M.E. García, and C. Cappo, editors, *XXXVI Conferencia Latinoamericana de Informática*. Centro Latinoamericano de Estudios en Informática, Facultad Politécnica – Universidad Nacional de Asunción and Universidad Autónoma de Asunción, Asunción, Paraguay, 2010. ISBN 978-99967-612-0-1.

Markines B., Cattuto C., Menczer F., Benz D., Hotho A., and Stumme G. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 641–650. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-487-4. doi:10.1145/1526709.1526796.

Menczer F., Pant G., and Srinivasan P. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)*, 4(4):378–419, 2004. ISSN 1533-5399. doi:10.1145/1031114.1031117.

Ponzoni I. *"Aplicación de Teoría de Grafos al Desarrollo de Algoritmos para Clasificación de Variables", Tesis Doctoral en Ciencias de la Computación, Directores: Simari G., Brignole N.B.* Ph.D. thesis, UNS, Bahía Blanca, 2001.

Schifanella R., Barrat A., Cattuto C., Markines B., and Menczer F. Folks in folksonomies: social link prediction from shared metadata. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 271–280. ACM, New York, NY, USA, 2010. ISBN 978-1-60558-889-6. doi:10.1145/1718487.1718521.

Tarjan R. Depth-first search and linear graph algorithms. *SIAM J. Comput.*, (1):146–160, 1972.

Xamena E., Brignole N.B., and Maguitman A.G. Computational models of relevance propagation in web directories. *Jornadas Argentinas de Informática e Investigación Operativa*, 40, 2011.