Graph-Based Domain-Specific Semantic Relatedness from Wikipedia

Armin Sajadi

Faculty of Computer Science, Dalhousie University, Halifax B3H 4R2, Canada
sajadi@cs.dal.ca
https://www.cs.dal.ca

Abstract. Human made ontologies and lexicons are promising resources for many text mining tasks in domain specific applications, but they do not exist for most domains. We study the suitability of Wikipedia as an alternative resource for ontologies regarding the Semantic Relatedness problem.

We focus on the biomedical domain because (1) high quality manually curated ontologies are available and (2) successful graph based methods have been proposed for semantic relatedness in this domain.

Because Wikipedia is not hierarchical and links do not convey defined semantic relationships, the same methods used on lexical resources (such as WordNet) cannot be applied here straightforwardly.

Our contributions are (1) Demonstrating that Wikipedia based methods outperform state of the art ontology based methods on most of the existing ontologies in the biomedical domain (2) Adapting and evaluating the effectiveness of a group of bibliometric methods of various degrees of sophistication on Wikipedia for the first time (3) Proposing a new graph-based method that is outperforming existing methods by considering some specific features of Wikipedia structure.

Keywords: Semantic Relatedness, Wikipedia Mining, Biomedical Domain.

1 Introduction

Measuring Semantic Relatedness is of great interest to philosophers and psychologists for its theoretic importance[7] as well as computer scientists for its various applications such as information retrieval[2]. Studies in both general domain and biomedical domain show that ontology-based methods outperform corpus based methods and are even more practical[4]. On the other hand lexical resources are expensive and labor-intensive, hence not available for most domains.

The first motivation for this project was assessing the suitability of Wikipedia as an alternative resource to domain specific human made ontologies for semantic relatedness problem by evaluating it in the biomedical domain, a domain with well known ontologies and successful graph based methods.

Wikipedia graph is not taxonomic, not even hierarchical. Most of the methods developed to work with knowledge bases are assuming this hierarchical structure

M. Sokolova and P. van Beek (Eds.): Canadian AI 2014, LNAI 8436, pp. 381–386, 2014.

and hence not applicable. On the other hand, this structure is more similar to citation graph extracted from scientific papers. The similarity between the two graphs was our second motivation to evaluate a class of similarity methods originally proposed for similarity in bibliometrics. We tuned and modified some of the methods to make them applicable on Wikipedia.

We also propose a new similarity method which considers a special feature of Wikipedia graph. Our experiments with bibliometric methods show that the increase in the usage of the graph structure results in a decrease in performance. We believe that it is a result of the dense structure of Wikipedia and its small diameter. Many nodes have more than thousands neighbors and in all of these methods, any neighbor plays the same role in calculating relatedness. There is more evidence backing up this claim [12] and this phenomena motivated us to propose a new algorithm which takes into account this fact by ordering the nodes by their role in the graph for calculating relatedness.

2 Methodology

A Wikipedia *page* is associated to each *concept*, so a directed graph $G_b(V_b, E_b)$ can be obtained with nodes (V_b) representing concepts and $(u, v) \in E_b$ if there is a text segment in the page associated with u pointing to the article associated with v. We call this *Basic Wikipedia Graph (G_b)*.

There are two types of edges, regular edges and *redirect* $E_r \subset E_b$ edges. Redirecting denotes synonymy. Let the *epsilon closure of a node* v, denoted by $\varepsilon(v)$, be the set of the nodes (concepts) accessible from v by travelling along redirect links. We use this concept to formally define the *Synonym Ring* of a node v, denoted by $sr(v)$, as the set of nodes equivalent to v. Finally the *Wikipedia Graph* can be formally defined by abstracting from this redirection using Synonym Ring. In this paper, by referring to a node associated with a concept v, we always mean $sr(v)$.

To compute the relatedness between concepts, we start from simple and well known graph-based methods. The class of methods we want to use are all based on this idea: "*Two objects are similar if they are related to similar objects*" [5].

2.1 Basic Bibliometrics and Simrank-Based Methods

Due to the similarity of the structure of Wikipedia to scientific papers, we propose evaluating well known bibliometrics on Wikipedia. Using bibliographic similarity terminology, we can count the portion of common incoming neighbors (*co-citation*), common outgoing neighbors (*coupling*), and combination of both through a weighted average (*amsler*).

The methods mentioned so far focus on only incoming or outgoing links and ignore the whole structure of the graph. If we want to go a step further and consider the relationships among the neighbors, one possibility is using a well known algorithm called SimRank[5]. If the rationale behind the recursion is generalized to outgoing links it is called *rvs-SimRank*[13] and if done for both

directions, it is called *P-Rank*[13]. Due to scalability limitations and the huge size of Wikipedia, SimRank is not runnable on demand (it runs globally), hence we do the recursions locally by defining the concept of *joint-neighborhood graph* for two nodes, which would be much a smaller graph.

2.2 Our Proposed Method: HITS Based Similarity

We believe that the unexpected decrease of performance with the increase of incorporating graph structure is a result of the density of the wikipedia graph and non-taxonomic relations: nodes mostly have high degrees (eg., USA has more than 500,000 neighbors) and therefor, not all edges have the same importance. For example, *September 2008* and *Clozapine* are both connected to *Schizophrenia*, while the former is just a date that among many other things, some new statistics about behavioural disorders was published, and the later is a drug to treat *Schizophrenia*.

Our idea is to rank the neighbors of each node based on the role they play in its neighborhood. We use Hyperlink-Induced Topic Search (HITS) [6], a well known concept in information retrieval originally developed to find authoritative sources in hyperlinked document set. We incorporate it in our class of similarity calculation methods, and refer to them by *HIT-Based methods* in this project. To find authoritative pages, it gives every node two scores: *hub score* and *authority score* and using the *mutual reinforcement* between the two, it iteratively updates them. So if it is run over a graph consisting of pages related to a concept (*focused graph* in HITS terminology), the final product of the algorithm is two ranked lists: authoritative pages and those which are good hubs to the authoritative pages.

Having two ordered lists, computing a real number showing the similarity between the two concepts can be done by comparing the two lists using Kendall's tau Distance[3]. Algorithm 1 is our proposed similarity method.

3 Evaluation and Conclusion

The standard datasets are a set of paired concepts with a human-assigned score which is considered to be the ground truth for their relatedness. The more scores reported by the automatic system correlate with the ground truth, the better the system is. The preferred method for this task is Spearman rank correlation (a.k.a Spearman's ρ). We base our evaluations on the most reliable dataset in biomedical domain, Pedersen et al.[11], known as Pedersen Benchmark. The dataset consists of 29 pairs scored by two different groups of physicians and medical coders. We compare our results to two reference reports, (1) McInnes et al.[8] (2) Garla et al.[4] which includes state of the art ontology based methods based on different ontologies such as SNOMED-CT, MeSH and *umls* (the aggregation of these two and 60 restriction free terminologies). Tables 1 and 2 reports the correlation between different automatic systems and human-scores.

From our experiments the following conclusions can be drawn:

Algorithm 1. HITS Based Similarity Computation

function $HITS\text{-}Sim(a,b)$
Input: : a,b, two concepts
Output: : Similarity between a and b
    $N[a] \leftarrow$ Extract a neighborhood graph for a
    $N[b] \leftarrow$ Extract a neighborhood graph for b
    $L[a] \leftarrow HITS(N[a])$
    $L[b] \leftarrow HITS(N[b])$
    $L'[a] \leftarrow append(L[a], reverse(L[b] \setminus L[a]))$
    $L'[b] \leftarrow append(L[b], reverse(L[a] \setminus L[b]))$
return $Kendall\text{-}Distance(L'[a], L'[b])$
end function
function $HITS(N)$
Input: : N, An adjacency matrix representing a graph
Output: : An order list of vertices
    $S \leftarrow$ Using HITS calculation, get hub or authority scores for each node
    $L \leftarrow$ sort vertices of N based on S
return L
end function

Table 1. Pedersen Benchmark: Comparison of correlations with[8]

	Physician		**Coder**	
	sct-umls	**msh-umls**	**sct-umls**	**mesh-umls**
path	0.35	0.49	0.5	0.58
Leacock & Chodorow	0.35	0.49	0.5	0.58
Wu & Palmer	-	0.45	-	0.53
Nguyen& Al-Mubaid	-	0.45	-	0.55
	Wikipedia		**Wikipedia**	
WLM	0.71		0.68	
co-citation	0.68		0.65	
coupling	0.66		0.63	
amsler	0.72		0.68	
SimRank	0.6		0.64	
rvs-SimRank	0.20		0.08	
P-Rank	0.39		0.41	
$HITS\text{-}sim_{aut}$	0.69		0.7	
$HITS\text{-}sim_{hub}$	0.69		0.63	
HITS-sim	**0.74**		**0.7**	

Table 2. Pedersen Benchmark: Comparison of correlations with[4]

	Physicians		Coders	
Ontology Based	Sct-umls	umls	Sct-uml	umls
Intrinsic IC*-Lin	0.41	0.72	0.49	0.70
Intrinsc IC-Path	0.35	0.69	0.45	0.69
Intrinsic IC-lch	0.35	0.69	0.45	0.69
PPR-Taxonomy relations	0.56	0.67	0.70	0.76
PPR-ALL relations	0.19	0.63	0.26	0.73
Wikipedia based	Physicians		Coders	
$HIT - SIM$	0.75		0.70	

*Information Content

1. Wikipedia based methods clearly outperform the most well known ontology based methods using different experiments. Only the combination of both SNOMED-CT, MeSH and other 60 terminologies can in some cases outperform Wikipedia based methods.
2. $HITS\text{-}Sim$ with Wikipedia gives the best results among all methods in most of the cases.
3. Our proposed HITS based methods and especially $HITS\text{-}Sim$ are performing well and give the best results in all cases.

4 Future Work

4.1 Theoretic Improvements

Although the idea of ranking neighbors is logical, the amount of improvement is less than our expectations. We believe that following modifications should shed more light on the problem:

- Decreasing the size of the neighborhood graph. Getting rid of even less important nodes by graph sampling algorithms has shown to improve the results of HITS and other ranking algorithms.
- Other ranking methods instead of HITS can be used straightforwardly, such as PageRank[1].
- Ranking neighbors to decrease the clutter around the nodes is not an alternative to the idea of SimRank which is propagation of the similarity. Hence we can use our idea to improve SimRank.

4.2 Further Evaluations

In this direction, we are interested in evaluating our similarity method on more datasets or in real applications.

- Automatic mapping of concepts to Wikipedia articles and disambiguation is a necessary module in case of dealing with large datasets. This task is more challenging in domain specific applications as the disambiguated sense should remain in the same domain.
- Experimenting with new datasets; although not many of such datasets exist, but recent studies[4] have used less reliable but bigger datasets (such as [9] and [10]).
- Applying our relatedness measure in other tasks, specifically *tag-generalization* and *Automatic indexing* of biomedical papers.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30(1-7), 107–117 (1998)
2. Budanitsky, A.: Lexical Semantic Relatedness and its Application in Natural Language Processing. Ph.D. thesis, University of Toronto, Toronto, Ontario (1999)
3. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2003, pp. 28–36. Society for Industrial and Applied Mathematics, Philadelphia (2003)
4. Garla, V., Brandt, C.: Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics 13(1), 1–13 (2012)
5. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 538–543. ACM, New York (2002)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46(5), 604–632 (1999)
7. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes 25(2-3), 259–284 (1998)
8. McInnes, B.T., Pedersen, T., Pakhomov, S.V.: UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In: AMIA Annu. Symp. Proc. 2009, pp. 431–435 (2009)
9. Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., Melton, G.B.: Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. In: AMIA Annu. Symp. Proc. 2010, pp. 572–576 (2010)
10. Pakhomov, S.V.S., Pedersen, T., McInnes, B., Melton, G.B., Ruggieri, A., Chute, C.G.: Towards a framework for developing semantic relatedness reference standards. J. of Biomedical Informatics 44(2), 251–265 (2011)
11. Pedersen, T., Pakhomov, S.V., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40(3), 288–299 (2007)
12. Yeh, E., Ramage, D., Manning, C.D., Agirre, E., Soroa, A.: Wikiwalk: random walks on wikipedia for semantic relatedness. In: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, TextGraphs-4, pp. 41–49. Association for Computational Linguistics, Stroudsburg (2009)
13. Zhao, P., Han, J., Sun, Y.: P-rank: a comprehensive structural similarity measure over information networks. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 553–562. ACM, New York (2009)