Comparing the Spread of Activation with the Nearest Neighbour Method in Semantic Email Search

Marcel Kvassay, Štefan Dlugolinský, Ladislav Hluchý

Institute of Informatics, Slovak Academy of Sciences,
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
{marcel.kvassay, stefan.dlugolinsky, hluchy.ui}@savba.sk

Abstract. Email search is a promising application area for semantic technologies with significant benefits for organizations and individuals with large email archives. Small and medium enterprises often conduct part of their business over email and their email archives contain a wealth of valuable information in semi-structured form. The ability to extract it and reason on it would be a distinct advantage. In this article, we compare the spreading activation algorithm with the nearest neighbour method on a simple reasoning task in the context of semantic search in a multidimensional social network extracted from an email archive.

Keywords: email, semantic search, spreading activation, nearest neighbour.

1 Introduction

Semantic search efforts currently focus mainly on Semantic Web, but there are other promising applications as well. One of them is the semantic email search that can benefit both individuals and organizations with large email archives. In our previous work [1, 2] we have focussed on small and medium enterprises that often conduct part of their business (e.g. sending and receiving orders or invoices) over email. Their email archives contain a wealth of valuable information in semi-structured form, such as customer names, phone numbers, postal addresses, products and their prices. We extract such “business objects” from emails with regular expressions and gazetteers, then place them as nodes into a multidimensional social network graph and reason on them with the spreading activation algorithm. One of the simple reasoning tasks that we used to test our approach was the assignment of the phone numbers to the people identified in the emails. Initially, the precision of our prototype was about 61% [1], but later we enhanced it so that its theoretical precision reached 85% [2]. We planned to improve it further by graph transformations, but our subsequent experiments in [3] showed that our algorithm was surprisingly resistant to changes in the graph on which it operated. Although the graph transformations reduced the number of nodes in the graph by 15 to 38 percent (by removing irrelevant nodes or nodes with low degree), the effect on the pairing precision was nil, or even negative. One possible explanation for this phenomenon was that in our specific conditions, the spreading activation might have actually “degenerated” into the nearest neighbour (shortest path) method, which, by definition, is not affected by the removal of irrelevant nodes. In this article we are primarily concerned with this hypothesis and with further improvement of our prototype as indicated by the results of our experiments.

© Jelínek J., Jiroušek R. (Eds.): Znalosti 2011, pp. 267–270, ISBN 978-80-248-2369-0.
VŠB–Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky, 2011.

2 Spreading Activation versus the Nearest Neighbour Method

Spreading activation can be described as a sequence of iterations (each consisting of a pulse generation and its spread over the network) that activates potentially relevant nodes in a semantic network [4]. There exist various spreading activation models, as well as different pulse computation techniques. We developed a simple breadth-first variant, which we applied to our multidimensional social network as shown in Fig. 1.

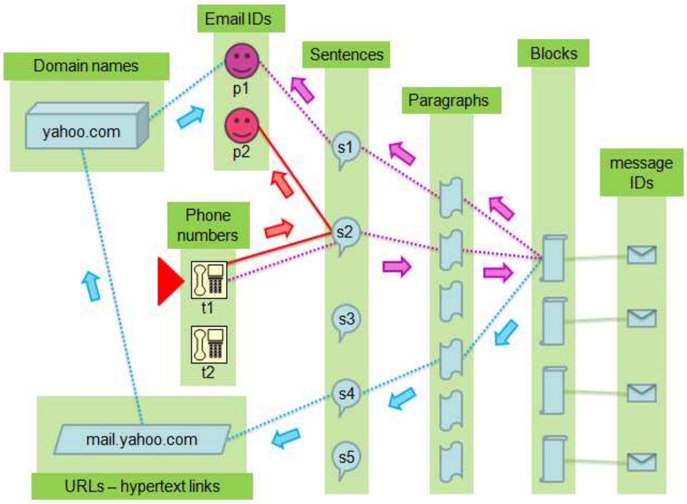


Fig. 1. Spread of activation in a multidimensional social network graph. Nodes represent actors (in this case, persons p1 and p2 identified by their email IDs), other "business objects" (especially the two telephone numbers t1 and t2 which we want to assign to their respective owners), and "structural" elements corresponding to the sentences, paragraphs and blocks of the email messages in which these objects were found. The solid red line connecting t1 with p2 through s2 means that both t1 and p2 were found in the sentence s2. In our case, this is also the shortest path from t1 to any person, so the nearest neighbour method would return p2 as the most probable owner of the phone number t1. In contrast, spreading activation takes into account also the longer (dotted) paths leading from t1 to p1 through the sentences s1 and s4. Though the activation gets attenuated each time it passes through an edge (so the shortest path carries over the greatest increment), the longer paths can be so numerous that their cumulative activation ultimately prevails. In such a case, the spreading activation would return p1 as the most probable owner of t1 in spite of the fact that p2 was closer to t1 in the graph.

Figure 1 demonstrates that the nearest neighbour method is considerably simpler than the spreading activation: instead of choosing the candidate with the maximum accumulated activation value, it just chooses the closest one. We implemented the nearest neighbour method mainly to see the difference in the pairing precision as compared to the spreading activation. If it were true that our prototype actually behaved like the nearest neighbour method, then there should be no difference. We used the same methodology as in [3], so the test task consisted in finding the "real owners" of the phone number, and the "real owners" were defined as the people who could be reached on that phone number. In this sense, one phone number could have several "real owners", e.g. all the people sharing the same office in a company.

3 Experimental Results

We present our results in the following tables. Table 1 compares the precision of the two methods in the classical way. The top two lines represent the spreading activation (SA) and the nearest neighbour (NN) methods searching the complete graph extracted from all the messages in the email archive. The two bottom lines represent their "localized" versions (LSA and LNN), where the graph for each phone number contained only the objects co-occurring with it in the same emails.

In general, for each phone number t_i, all the methods return a list of candidate owners [c_{i1}, c_{i2}, c_{i3}...] sorted in descending order either by their shortest path ranks, or by their accumulated activation scores. The pairing is accepted as correct when a "real owner" occurs as the first or the second in the candidate list. In fact, both the NN and LNN methods also use the activation scores in a subordinate role – in order to distinguish among the candidates with the same shortest path rank.

Table 1. Precision of the spreading activation (SA) and the nearest neighbour (NN) methods

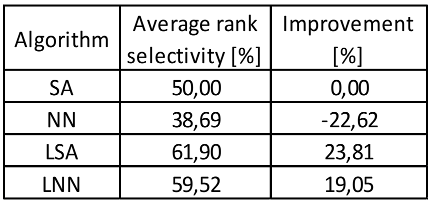
Algorithm	Phones Total	Correctly paired	Precision [%]
SA	24	19	79,17
NN	24	16	66,67
LSA	24	22	91,67
LNN	24	22	91,67

The results show that SA significantly outperformed NN in the complete graph search, which means the resistance to graph changes that we had observed in [3] was not caused by SA "degenerating" into NN, but represented an intrinsic and valuable robustness of the spreading activation algorithm itself.

In the localized search, both methods seemed to perform equally well. We felt there might be hidden differences, but we needed a more fine-grained measure to visualise them. We could not reuse the measure that we had defined earlier in [3], since the shortest path ranks were not directly comparable with the spreading activation scores. In the present study we therefore assigned simple "point" scores to the top three candidates in the list: the topmost candidate was given 4 points, the second one 2 points, and the third one 1 point. The other candidates were assigned 0 points. We defined a new measure for each pairing task – "rank selectivity" (*RS*) – as the score accumulated by the "real owners", divided by the theoretical maximum of 7 points. The average rank selectivity (*ARS*) for each method was then defined as the average of the rank selectivity scores for all the paired phones.

Table 2 shows the *ARS* values for all the four method variations, and the percentage improvement that they represent over the spreading activation in the complete graph (SA), which served as a baseline. The results show a small but noticeable difference between the two methods even in the "localized" mode. Both methods are clearly correlated (since in most cases the closest person is indeed one of the "real owners" of a given phone number), yet the spreading activation method seems to be the better and more robust of the two, as it suffers less from the imperfections of the underlying social network graph.

Table 2. Average rank selectivity for the compared methods



Algorithm	Average rank selectivity [%]	Improvement [%]
SA	50,00	0,00
NN	38,69	-22,62
LSA	61,90	23,81
LNN	59,52	19,05

4 Conclusion

We intend to test the statistical significance of our results and further analyze the spread of activation in social networks. Search localization can be considered a special type of graph transformation with one parameter – the attribute instance (phone number) which is to be assigned to a primary entity (owner). With localization, our prototype crossed the 90% level of precision, which we take for a sign of its approaching maturity. It means we can subject it to more demanding tests in future, e.g. to require that the correct owner of a given attribute instance comes strictly first in the candidate list. We see promising opportunities for further improvement, especially if we extend our social network graphs with new structural elements (e.g. new edge types or edge attributes) and adapt our spreading activation algorithm to exploit these new kinds of information.

Acknowledgments

This work is partially supported by the projects RECLER ITMS: 26240220029, SMART II ITMS: 26240120029, AIIA APVV-0216-07, VEGA No. 2/0211/09 and VEGA 2/0184/10. We would also like to thank Fedit for providing us with emails for testing.

References

1. Kvassay, M., Laclavík, M., Dlugolinský, Š.: “Reconstructing Social Networks from Emails”. In: Pokorný, J., Snášel, V., Richta, K. (eds.): DATESO 2010: *Proceedings of the 10th annual workshop*. MATFYZPRESS publishing house, Faculty of Mathematics and Physics, Charles University, Prague (2010) 50-59. ISBN 978-80-7378-116-3
2. Michal Laclavík, Marcel Kvassay, Štefan Dlugolinský, Ladislav Hluchý: Use of Email Social Networks for Enterprise Benefit; In: International Workshop on Computational Social Networks (IWCSN 2010), IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010
3. Kvassay, M., Laclavík, M., Dlugolinský, Š., Hluchý, L. (2010): Graph Transformations for Semantic Email Search. In: 5th Workshop on Intelligent and Knowledge Oriented Technologies : WIKT 2010 proceedings. - Bratislava : Ústav informatiky SAV, 2010, pp. 64-67. ISBN 978-80-970145-2-0
4. Crestani, F.: " Application of Spreading Activation Techniques in Information Retrieval". (URL: http://www.springerlink.com/index/g11t185158667418.pdf)