

Mining Socio-Semantic Networks Using Spreading Activation Technique¹

Alexander Trousov

(IBM Dublin Software Lab, Ireland
atrousso@ie.ibm.com)

Mikhail Sogrin

(IBM Dublin Software Lab, Ireland
sogrimik@ie.ibm.com)

John Judge

(IBM Dublin Software Lab, Ireland
johnjudge@ie.ibm.com)

Dmitri Botvich

(Waterford Institute of Technology, TSSG, Ireland
dobtvich@ie.tcg.org)

Abstract: A mining method for egocentric and polycentric queries in multi-dimensional networks is proposed. The method allows fast search for objects in sufficient proximity of other object(s) where the proximity is defined in terms of multiple relationships between objects. The method uses spreading activation technique. Other potential uses of spreading activation technique are also outlined and, in particular, include applications to collaborative filtering (community detection based on tag recommendations, expertise location, etc). Moreover, the spreading activation technique is combined with so-called ambient navigation. The advantages of such approach are high performance and high scalability in terms of size of multi-dimensional network. The proposed method is very practical and is implemented in IBM LanguageWare software products.

Keywords: Spreading activation, social web, semantic web

Categories: H.3.1, H.4.1, H.5.2, H.5.3, I.2.7, I.5.3, L.1.3, L.1.4, L.6.1, L.6.2

1 Introduction

The proliferation of Web 2.0 has brought together people and all kinds of digital artefacts: documents, concepts, vocabulary, tasks, activities, and more. The Web is increasingly becoming a participatory, social space which has established notions such as tagging as a popular mechanism to replace hierarchical categorization and formal ontologies.

A type of Socio-Semantic Web is emerging. It focuses on personalization, collaboration, findability and navigation, and bottom-up conceptualization. This web

¹ This material is based upon works supported by the European Commission under the Nepomuk project FP6-027705.

is a multi-dimensional network which connects people and the things they create and do. Each dimension of this emerging web represents different types of data with different types of relationships. Semantic Web techniques which allow us to specify types of these links, and how to use them are also suited for use in this field.

In this paper we'll discuss how technique of spreading activation helps in creating solutions for these types of multi-dimensional networks (people, documents, tasks, etc.). We'll illustrate this using as an example some new technology based on spreading activation developed by IBM LanguageWare which provides an integrated platform for combining social computing, semantic processing, and activity-centric computing for enhanced user experience.

This paper is organised as follows: in Section 2 we describe the spreading activation technique, Section 3 describes the initial input, and Section 4 describes how we can use the results and "tune" for specific tasks.

2 Spreading Activation Technique

Cognitive psychology and artificial intelligence research model reasoning and memory as processes on neural networks. These networks of neurons and the patterns in which they fire simulates certain aspects of the human brain. There are many different algorithms and implementations which model these processes, one of which, spreading activation, we use to mine information from multi-dimensional network data such as socio-semantic networks [Anderson,83].

Our approach focuses on using multi-dimensional networks comprised of concepts which represent actors (people, organisation, etc) and various digital artefacts related to the things they create and do, these concepts (nodes) are linked together by various socio-semantic links which represent relationships between concepts. Nodes and links in these networks may be typed and weighted, which allows great expressivity and applicability to various data and scenarios.

In general, the spreading activation algorithm proceeds as follows:

1. Initial activation is set to one or several nodes in the network (e.g. with value 1.0). This initial activation may represent items of interest, context of a document, user profile, etc.
2. Activation is spread to neighbouring nodes, but the activation value is normally less than the value of a source. For this, an activation decay parameter is introduced, usually in the range [0...1]. As the activation spreads through the network, different link types may have associated different decay values allowing for different effects like a lower rate of decay through "preferred" links.
3. If activation is spread from a node with many links, those neighbouring nodes will get even less activation to simulate a situation that many similar items get less attention when compared to one unique item.
4. However, if there are multiple paths in the network to some node, its activation will be sum of activations from its inputs. And therefore, it may get activation value even higher than the source.
5. After all activation values are calculated, they are ranked and nodes with higher activation represent important or interesting items or concepts.

We implemented a spreading activation algorithm in our Galaxy application, available from the IBM AlphaWorks site². It contains a Java library for spreading activation and graph mining, as well as a GUI application based on Eclipse RCP.

Initial and final activation can be seen as functions on the network (e.g. real-valued function on nodes of the network), and spreading activation algorithms as a transformation from initial to final functions. It also allows us to “smooth” the initial activation function and get “highlights” or areas where activation peaks.

For example, in [Kinsella,07], we demonstrated that initial activation on four corners of a small grid might detect the centre of the grid. Some other examples will be presented later in this paper.

Superficially, the Galaxy UI appears like Google Sets³: the user focuses on one or more concepts and Galaxy tries to predict other concepts of interest (by returning the ranked list of the most activated nodes). However, unlike Google Sets, Galaxy can return results which are not necessarily in the same set of objects as the input, but which are still related. For example, given the input “red,” “green,” “orange” Galaxy might return a list of other colours, but it can also return concepts which are related, but not necessarily colours, such as “traffic lights.”

3 Processing egocentric and polycentric queries

Spread-of-activation might be used to provide rapid egocentric and polycentric search: query for search is expressed by placing initial activation in the nodes of interest; the result is the list of other nodes on network, ranked according to the cumulative strength of their connection to the initially activated set of nodes.

Here, we provide illustrative numerical simulations for typical small graph clusters. The nature of the spreading activation implementation we use means that the results in large scale graphs which have similar clusters contained within them as sub graphs will be similar. In the section 3.1 we provide considerations regarding cognitive value of the results of egocentric queries. In the section 3.2 we discuss applications of polycentric search. These results might be used for search and navigation in the multi-dimensional networks.

3.1 Egocentric Querying

By providing activation to a single node of interest in the network and allowing the activation to propagate through the network the result is a low level analysis of what concepts (nodes) are closely related to the node of interest. This analysis is derived based on the existing relationship links in the network and the cumulative strength of activation spread through these various links. The objective of such egocentric queries is to derive an overview of the most relevant available content relating to a particular node or concept based on the data encoded in the semantic network.

[Kinsella,07] describes how egocentric queries on a social network can be used to find areas of interest, people and documents relating to individuals but which are not explicitly linked to that individual. Here, we approach egocentric queries from a

² <http://alphaworks.ibm.com/tech/galaxy>

³ <http://labs.google.com/sets>

different point of view and show how we can use this approach to acquire additional knowledge from the network about the topic of interest of the egocentric query.

To illustrate this we use a rather simple example to show how the process works. In our example the socio-semantic network contains a person who is connected to some other people, who are researchers. Depending on the configuration of the network we would like to be able to use an egocentric query with Galaxy to extract the information that our person of interest is connected to other people or simply to researchers. Figure 1 illustrates this scenario.

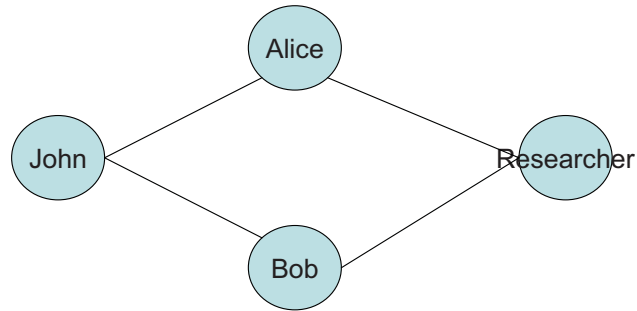


Figure 1: John is connected to two people who are researchers

Given a network like that in Figure 1 if we perform an egocentric query on the node labelled “John” it would not be unexpected that Galaxy would return “researcher” as a node which is related to John, but it would be unexpected for “researcher” to be favoured over the other two nodes in the network. That is to say that based on the information encoded in the network that we would not expect a response to our egocentric query to generalise and favour saying that “John is connected to researchers” over saying “John is connected to Alice” or “John is connected to Bob.” Figure 2 shows the activation values returned when we query such a network with Galaxy.

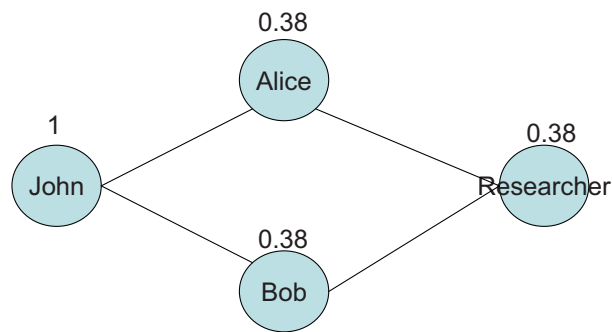


Figure 2: Activation values after egocentric query for “John” which was expressed by placing initial activation with the value 1.0 in the node “John”

The activation values shown in Figure 2 show that Galaxy has performed an accurate analysis of the network and determined that a relationship exists between “John” and his immediate neighbours in the network, that an indirect relationship exists with the concept “researcher” and that none of these relationships can be deemed to be stronger than the others based on the data in the network. This is a somewhat simplified example, however if we expand the network by introducing some more nodes as in Figure 3 we can make some interesting observations.

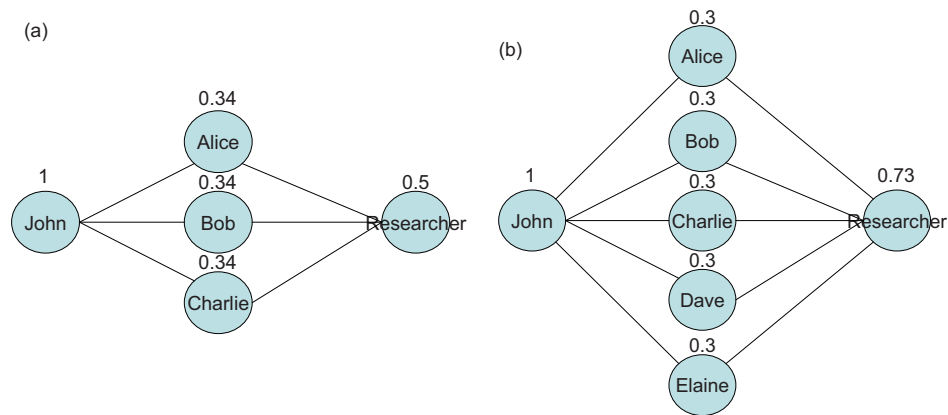


Figure 3: Egocentric query on larger networks

The graphs in Figure 3 show that as the number of intermediary nodes increases the activation level in the general category “researcher” increases while that in the individual nodes for each person is only slightly reduced. This is a desirable result because under these circumstances the nodes linked to “John” can be generalised more completely by the concept “researchers.”

3.2 Polycentric queries

An extension of simple egocentric queries, polycentric queries are those where initial activation is provided to multiple nodes in the network. These type of queries allow us to discover a wider range of new facts from the network data. The activation spreading from multiple simultaneous egocentric queries combines and accumulates in certain nodes which can increase the overall ranking of some nodes which would previously have had very little activation, and so would have been ranked lower. In this way by activating two or more nodes at the same time we can find nodes representing people, interests, documents etc., which are directly and indirectly related to both concepts and then use their relative activation levels to rank them according to their relevance to the overall polycentric query instead of to an individual egocentric query. This allows us, for example, to find people who are related to a series of documents, or to locate experts on a given topic who are in an individual’s social network.

4 Ambient Navigation

Today's applications which provide access to information typically use two navigation methods, *browsing* and *search*. A browser provides a number of links to other items or documents, while a search system shows relevant items given user's query.

We introduce a new navigation method in the socio-semantic space named *ambient navigation*, which utilises semantic network and user's context. It provides the convenience of having items presented in a way tailored to a particular user, situation, activity or context, but retaining a freedom of browsing.

Ambient navigation in our implementation is on-the fly transformation of the underlying semantic network (or ontology) in such a way that users can focus on one or more nodes (concepts) in the network, and immediately see a conceptual summary of their focus, in the form of transformed reduced network, in which unrelated items will be pruned (but not removed completely) and highly relevant items will be brought to the user's attention even if they were not explicitly linked to current user's focus. So users can navigate the information in a guided yet unconstrained way.

Ambient navigation is our user-centric generalisation of the theoretical approach which some research calls "dynamic taxonomies" [Sacco,00], but we don't restrict ourselves to hierarchical structures and facets. This kind of navigation is fast and easier for users to perceive, contextualise, simplify, and make sense of otherwise complex interlinked data without cognitive load.

4.1 Applicability of ambient navigation to collaborative tagging systems

Collaborative tagging or bookmarking systems like Del.icio.us or IBM's dogear are currently growing in popularity. They allow users to bookmark and tag web resources as well as view other people's bookmarked resources and tags.

All the data describing a given instantiation of a collaborative tagging system can be viewed as networks, where users, resources and tags are related by instances of tagging. Using this network, Galaxy can be used for a variety of scenarios:

- *Community detection*: Galaxy can find a community of people who relate to some other people when they tag similar resources with similar tags, therefore having similar interests or expertise. But community detection can also be used with regard to specific web resources or tags, or even any combination of them. When community detection is performed based on the current focus of a tagging system's user, it represents an example of ambient navigation by presenting useful contextual information and guiding a user to something that may be interesting for him.
- *Tag recommendation*: When tagging a web resource, a user may be presented with recommended tags based on his community. It would not simply present most frequent tags, but select ones that are used for that resource by people in his community.

The aim of ambient navigation is not to make choices for the user, but to improve or speed up user's navigation decisions by providing the most probable answers and links which are most likely to be of interest.

All these are applications of polycentric queries, as an initial activation may be put into multiple items like resources and tags. For instance, social metadata (tag) recommendation may proceed as follows:

1. Put initial activation at a person and a resource, and then find other related people or community.
2. Put initial activation at the same resource and people from the community found in the previous step, and find what tags would be related to that resource.

In this example spreading activation procedure was “cascaded”, i.e. used several times with the results of a given run used as a basis for subsequent ones. In addition, intermediate results of calculations may be presented to the user in some form to provide a means of explanation as to why and how the given results were achieved.

5 Conclusions

Spread of activation is a useful technique which can be easily used to mine generic or specific data represented by networks such as social networks and the semantic web, or the emerging socio-semantic web. Our implementation can be optimised for a particular task using parameters for single acts of spread-of-activation, or through cascading the results of several activation passes. In addition to this, its relative simplicity makes it computationally efficient, scalable, and extensible.

Our implementation gives results which reflect the context accurately and which are based on the context of the query and the underlying socio-semantic network and which can be tuned to variety of tasks including social/collaborative bookmarking, community detection and expertise location, as well as many emerging tasks.

6 Future Work

We believe that solution to many tasks in socio-semantic can be viewed as navigation from one or more initial nodes to other sets of nodes. Spreading activation techniques are generic, extensible, scalable and applicable to a variety of those problems (especially as ambient navigation).

The downside of this generality is that because the algorithm is optimised for speed, tuning can be hard, even using parameters which are described in [Galaxy,07]. We plan to address this in our future work. Our set of parameters is explicit and our performance is fast, so there is great potential for supervised learning to optimise the parameters. Randomised search in the parameter space (e.g. using evolutionary computing techniques) could prove to be a promising solution.

Secondly, we have illustrated how simple cascading of the results of spreading activation improves robustness and allows the user to easily guide the process. There is a wealth of possibilities for developing applications using this method once we have established an effective means to tune the base parameters for each task.

References

[Anderson,83] Anderson, J.: A Spreading Activation Theory of Memory, *Journal of Verbal learning and Verbal Behavior* 1983, (22), 261-295

[Kinsella,07] Kinsella, S., Harth, A., Trousov, A., Sogrin, M., Judge, J., Hayes, C., Breslin, J., Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects, In Proc. In Proceedings of Applications of Social Network Analysis ASNA 2007, September 2007

[Galaxy,07] IBM LanguageWare Miner for Multidimensional Socio-Semantic Networks, December 2007, <http://www.alphaworks.ibm.com/tech/galaxy>

[Sacco,00] Sacco, G.: Dynamic taxonomies: a model for large information bases, *Knowledge and Data Engineering, IEEE Transactions on* Volume 12, Issue 3, May/Jun 2000, 468 - 479