Galaxy: IBM Ontological Network Miner

John Judge, Mikhail Sogrin, Alexander Troussov

{johnjudge, sogrimik, atrousso}@ie.ibm.com

Abstract: Many applications of the semantic web and Web 2.0 aim to empower the knowledge worker. These applications however, do not allow the user to combine all of his/her social and semantic information into a single resource which allows data to be processed, managed and enhanced automatically. In our demo we will present a number of demo applications based on Galaxy, IBM's ontological network miner, which was designed to work with such resources to enhance the capabilities of a number of applications in social semantic computing. Galaxy is a highly efficient, scalable system which can be easily built into an application and can be optimised to suit a user's preferences or to take into account the needs of a particular task or application.

1 Introduction

Currently the semantic web relies on semantic annotations which, for the most part, are done manually by humans. Working in the EU 6^{th} framework project Nepomuk [Nepo] we in IBM Dublin have developed a tool which can be useful in the automation of metadata creation. Our ontological network miner, Galaxy, is a generic tool which performs elements of soft clustering on semantic networks such as company organisation trees, social networks and community diagrams or any other collection of data which can be represented by a graph network.

We perform automatic ontology-based conceptual tagging and find central concepts of a text with respect to the given lexico-semantic resource (ontology). For example, a text which mentions Mulhuddart, Lansdowne, Clontarf is probably about Dublin/Ireland/Europe/Earth. This fact can be inferred (assuming some geographical information exists in our semantic resource) from geographical relations like `Mulhuddart is-part-of Dublin`. Galaxy resolves any ambiguities on the fly based on the ontological knowledge from the corresponding semantic resource and uses the results of disambiguation in determining the results.

This kind of processing can be leveraged for numerous tasks including metadata generation, related item recommendation, community detection, and expert location. We have designed our application in such a way that it is highly configurable to make it adaptable to numerous tasks in social semantic computing.

The remaining sections of this paper are structured as follows: Section 3 describes our network mining algorithm and gives some performance statistics. Section 4 discusses some of the many applications which our algorithm could be used for. Finally Section 5

outlines some future directions for our work.

2 Motivation

[Tof90] observes that knowledge workers in the age of knowledge economies and knowledge societies need to have available to them a system which can be used to create, process, enhance and manage their knowledge and information. Recent advances in social computing and social semantic desktop applications are making such a system possible. However, many of the resources available for these tasks need significant manual intervention before they are useful to the knowledge worker. We have created a highly scalable and efficient ontology mining algorithm which can be used for a variety of tasks in social semantic computing and which can be developed into an application which can suggest links between resources to remove the need for manual intervention and which can be adapted to a user's preferences or for individual tasks.

The Nepomuk project aims to empower knowledge workers to better exploit their personal information space and to maintain fruitful communication and exchange within social networks irrespective of organizational boundaries. In the context of Nepomuk we are working with our partners to develop a comprehensive solution which extends the personal desktop to create a collaboration environment which supports both personal information creation, processing and management, and the sharing and exchange of information across social and organizational relations. This solution is called the Social Semantic Desktop (SSD).

The SSD is built upon the idea of a Personal Information Management Ontology (PIMO) which is a unified model of social and semantic data. The PIMO is neither a fixed nor a hierarchical entity, it grows and changes as the user creates new data, uses existing data and changes his/her social interactions. Because of the organic and dynamic nature of the PIMO an efficient, scalable method of mining information from the ontology and of inferring new data based on the topology is required to exploit this data fully.

Much of the existing network mining technology is lacking in this regard, often they rely on a rigid or hierarchical ontology structure or they suffer badly in terms of complexity on large datasets. For example [AHSS04] presents an accurate scalable algorithm which assigns a geographical focus to a texts based on mentions of places in the text. However their method requires that the underlying datasource is structured hierarchically and it is confined to just one domain of application. Galaxy is a significant improvement on these type of miners, not only is it efficient and scalable but because it makes no assumption about graph topology it can work on an ontology of any complexity.

3 Description

Galaxy is based on the spread of activation technique used in semantic networks. It is very flexible and customisable which allows it to be tailored for a range of applications. The spread of activation is used to find a focal node, or nodes, in the network based on the parameters and constraints given. Galaxy resolves any lexico-semantic ambiguities on the fly based on the ontological knowledge from the corresponding resource and uses the results of disambiguation in determining the focus.

Galaxy does not perform clustering in the traditional sense ("hard clustering"), where a graph or network is partitioned according to various clustering measures. Instead Galaxy performs a "fuzzy clustering" analysis, dealing with a (changing) sub-graph based around a set of nodes within the graph provided to it, and finds a focus (or foci) relative to those nodes and dependent upon the graph topology and the user's constraints on how propagation around the graph can happen. The focus found by Galaxy is similar to finding a central node or concept for the given sub-graph. However, dependent on the starting nodes, the constraints of the specific application and graph topology, multiple foci or no focus may be returned. In this way it does not return a central concept or focus unless one can be found which is close to the starting nodes.

Testing Galaxy as a stand alone entity is somewhat difficult. Standard metrics like precision and recall are difficult to apply without a particular task in mind and a specific test set and lexico-semantic resource for that task. Galaxy would also have to be incorporated into a "driver application" which would perform the particular task. This means that as yet we are unable to supply qualitative results on Galaxy's performance.

We have performed scalability tests on Galaxy to test how well the algorithm copes with different amounts of nodes activated in the initial query. The tests were carried out on a network of over 170,000 nodes, up to 100 initial nodes were chosen at random to activate and the time taken for each query length were averaged over 10 runs.

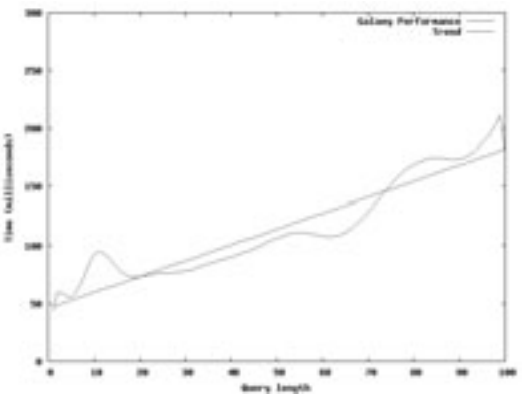

Figure 1: Graph of time versus query length performance for Galaxy

The graph in Figure 1 shows that the time taken to execute a query increases as the number of activated nodes in the query is increased. The overall trend in the time increase is roughly linear. The time taken to execute individual queries is remarkably small for such a large network, and even when the longest queries tested can execute in the region of 0.2 of a second.

4 Applications and the Demo

Galaxy can be used to add value to a range of different applications. We have developed a prototype application "workbench" which allows us to demonstrate a number of tasks useful for IBM's new enterprise corporate social software solution Lotus Connections [LSS]. We will demonstrate ways in which Galaxy can tie together resources from social software at IBM for a number of tasks including metadata generation, tag recommendation, community detection and expertise location.

Our demo will feature live demonstrations of these tasks and more applications as we discover them and refine our demonstration workbench. We will also demonstrate a composite application built on Lotus Notes 8 called Smart Assistant, which uses text analytics and Galaxy to organise incoming email and provide contextual information based on mail content and real world knowledge in the form of a semantic ontology.

5 Future Directions

As yet Galaxy is in the early stages of being deployed into real world situations and applications. We are working closely with our partners in Nepomuk, Digital Enterprises Research Institute Galway, and Trinity College Dublin to explore new areas of application and to build solutions based on Galaxy. This is an ongoing challenge and many new areas of research are open to us including using Galaxy as an application development component instead of just a monolithic ontology mining algorithm.

References

[AHSS04] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: Geotagging Web Content. In *SIGIR*, pages 273–280, 2004.

[LSS] Lotus Social Software http://www.ibm.com/lotus/connections

[Nepo] Nepomuk - The Social Semantic Desktop http://nepomuk.semanticdesktop.org

[Tof90] Alvin Toffler. *Power Shift, Knowledge, Wealth and Violence at the Edge of the 21st Century*. Bantam Books, 1990.