Knowledge Media Institute

A Spreading Activation Framework for Ontology-enhanced Adaptive Information Access

Md Maruf Hasan, Enrico Motta, John B. Domingue, Simon Buckingham-Shum, Maria Vargas-Vera, Mattia Lanzoni

KMI-TR-122

September, 2002

www.kmi.open.ac.uk/tr/papers/kmi-tr-122.pdf

A Spreading Activation Framework for Ontology-enhanced Adaptive Information Access

Md Maruf Hasan, Enrico Motta, John B. Domingue, Simon Buckingham-Shum, Maria Vargas-Vera and Matia Lanzoni

Knowledge Media Institute

The Open University, Milton Keynes, MK7 6AA, United Kingdom

{m.m.hasan, e.motta, j.b.domingue, s.buckingham.sham, m.vargas-vera, m.lanzoni} @open.ac.uk

Abstract

This research investigates a unique Indexing Structure and Navigational Interface which integrates (1) ontology-driven knowledge-base (2) statistically derived indexing parameters, and (3) experts' feedback into a single Spreading Activation Framework to harness knowledge from heterogeneous knowledge assets. Within an organisation, organisational ontologies capture precise knowledge about organisational entities: people, projects, activities, information sources and so on. We extract useful entities and their relationships from an ontology-driven knowledge base. We also process collections of documents (archives) accumulated in heterogeneous information-bases within an organisation and derive indexing parameters. This information is then mapped to a weighted graph (spreading activation network). The network contains three distinct sets of nodes representing documents, ontological entities and statistically derived entities. Document nodes are connected to both ontology-driven entities and statistically derived entities, and vice-versa with relevant weights. Retrieval is performed by spreading query-based activation into the network and selecting the most-activated nodes. Experts as well as users in the organisation either navigate the network using associative relations among nodes or with specific queries. Expert's feedback is captured and the network weights are continuously adapted. This framework essentially combines precise knowledge (ontology-driven), non-precise knowledge (statistically driven) and Expert's feedback (adaptation and refining) into a single framework for effective information retrieval and navigation.

Introduction

In recent days, it is predicted that within a decade or so, computing technology will transform the Internet into the Interspace (Schatz, 2002), an information infrastructure that supports semantic indexing and concept navigation across distributed and non-distributed community repositories. In the past, some Artificial Intelligence (AI) researchers have moved from a concern with manually constructed knowledge representation to automatic Machine Learning. With the advent of the Semantic Web, more and more AI researchers have been revisiting knowledge representation issues – for instance, the renewed interest in ontology (Lopatenko, 2001; McGuinness, 1998, 2002; Middleton et al. 2002). The Information Retrieval (IR) community has also begun to move from fully automatic indexing approaches to partially manually crafted approaches - for instance, adaptation and personalisation of search and retrieval by capturing how indexing structures change with use. The recent trend is that the two methodologies will be increasingly overlapping and applied together to solve practical problems (Belew, 2000).

Organisational ontologies codify precise information about the organisation: its people, projects, core activities and information-bases. Useful features such as, organisational entities and their relationships are easily extracted from the ontology-driven knowledge base. Archives of documents accumulated in the organisational information-bases over time also contain implicit and explicit knowledge about the organisation. By analysing a collection of documents using statistical and natural language processing techniques, we also identify useful indexing units and annotate some of these units successfully with appropriate labels (e.g., people, projects, etc.). In this paper, we investigate a bootstrapping approach between effectively using the handcrafted precise knowledge (from the ontology-driven knowledge base) and automatically extracted non-precise knowledge (from the archives) to enhance the information access experience in an organisational setting.

A Spreading Activation network (Crestani, 1997, 2000) is created for all the documents using indexing units (statistically derived) and entities (ontology-driven). The central nodes are the documents. Document nodes are connected to and from these auxiliary nodes consisting of both precise and non-precise items with appropriate weights. The spreading activation dynamics are maintained by putting a constraint on the network - the sum of weights of all outgoing edges from each node is constant at all times. Users may navigate the network through associative links as well as by sending queries. Queries can be progressive and the relevance feedback can also be fed into the network. Retrieval is performed by spreading query-based activation into the network and selecting the most-activated nodes.

The Organisational Scenario

Organisational knowledge is implicitly or explicitly captured in written documents (manuals, publications, news releases, etc.). Such documents are typically organised under different servers. It is common within organisations to have a number of archives and databases and their associated services. The following Figure (Figure 1) depicts the AKT (Advanced Knowledge Technology) scenario, where there are a number of document servers: the *AKT Planet Server* (http://news.kmi.open.ac.uk/akt/), the *AKT E-Prints Server* (http://eprints.aktors.org/) and so on.

The AKT Consortium (http://www.aktors.org/) is a long-term collaboration among 5 UK universities, which has been investigating different issues with knowledge management. The AKT consortium has its own *Reference Ontology* (http://kmi.open.ac.uk/projects/akt/d3e/2002/ref-onto.html) deployed on the *WebOnto* server (http://webonto.open.ac.uk/, Domingue et al., 1999). The AKT reference ontology codifies the academic life in terms of projects, people, organisations and their interrelationships. The ontology-driven knowledge base captures organisational entities and their relationships and therefore facilitates intelligent inferences about different entities in the organisational context. Such inferences are very helpful in IR context, for instance, query expansions. AKT related activities and news are posted regularly on the AKT Planet Server (http://news.kmi.open.ac.uk/akt/). Publications made by different individuals and groups are being submitted continuously to the AKT E-Prints Server (http://eprints.aktors.org/). There are also many other sources of structured and unstructured information within the AKT consortium: AKT Internet and Intranet sites, discussion forums and portals. We can enhance organisational information access to a greater extent when we integrate heterogeneous sources of information and bootstrap them together. For example, author-assigned metadata available for each document in the E-Prints archive may correlate with documents in Planet news archive where there is no explicit metadata available for the news stories.

Information retrieval using full-text search is almost always readily available for any collection of documents. However, full-text search and retrieval are ambiguous since they work on content (i.e., representation) level. Some contextual (or semantic) enhancements in accessing information can be achieved at the cost of manual (or semi-automatic) assignment of metadata and taxonomic information. Documents in the E-Prints archive take advantage of such metadata. However, AKT-Planet stories have no such metadata. In the AKT context, the metadata defined in AKT E-prints may also be relevant with the AKT-Planet news stories. Moreover, having the AKT-ontology up and running, we know the big picture of the organisational context (the relevant entities and their relationships). The indexing structure used in this research is based on bootstrapping knowledge acquired over heterogeneous sources within an organisational context.

Like any organisation, as the AKT project grows, the number of archives continues to grow and so does the number of items within each archive. The AKT scenario is dynamic, as new people are likely to join (or leave) the project, new technologies are likely to be developed and so on. An integrated and effective way of accessing and processing organisational information and knowledge from all these heterogeneous repositories is therefore inevitable.

The AKT consortium provides the organisational scenario in which we address effective information access issues.

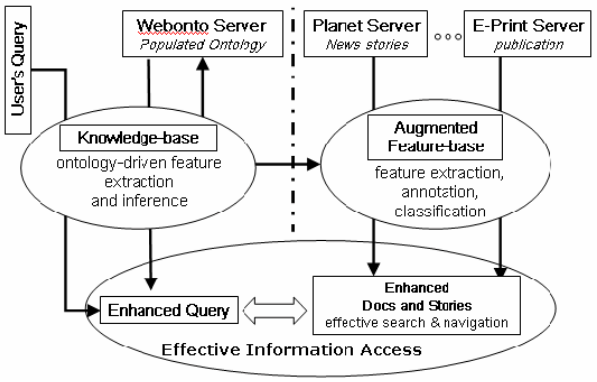

Fig. 1: AKT Scenario

Indexing and Navigation Strategies

Augmentation rather than automation of human performance is technically feasible for an ontology-enhanced information access environment. Such augmentations can be achieved from the ontology-driven knowledge base as well as automatic processing of a collection of documents.

Based on the observations in the ontology-driven knowledge base and the metadata, we use a machine learning algorithm (Frank et al. 1999) to extract and annotate known and unknown entities in the documents. We call this process *Entity Recognition and Annotation*. Similar to full-text indexing, we also compute statistical parameters (term-frequencies, inverse document frequencies, mutual information) from the collection of documents to identify non-trivial indexing units. Each document is then represented as a vector of entities (ontology driven) and indexing-units (statistically derived).

Basic Representation: Each new document first causes a corresponding document node to be generated. Nodes of

ontological entities (like people, project and technologies etc.) in the document are then created (if they do not already exist) and subsequently connected with the document node. Two links are created between the document and each of its entities, one in each direction. Similarly, nodes are created for each indexing unit and connected to the corresponding documents in the same fashion. Weights are assigned to these links according to an inverse frequency weighting scheme (Salton, 1988). The sum of the weights on all links going out of a node is forced to be a constant to maintain spreading activation dynamics.

The initial network is constructed from the superposition of many such documents' representations (c.f. Figure 2). A network built using this scheme, has the satisfying property of conserving activity. That is, if a unit of activity is put into a node and the total outgoing associativity from that node is constant, the amount of activity in the system will neither increase nor diminish. This is helpful in controlling the spreading activation dynamics of the network during querying.

Querying and Retrieval: Users may begin a session by describing their information need, using a simple query language. An initial query is composed of one or more clauses. Each clause can refer to one of the types of "features" represented in the network: free vocabulary, people, projects, technologies and other known entities. This query causes activity to be placed on nodes in the entire network corresponding to the features named in the query. This activity is allowed to propagate throughout the network, and the system's response is the set of nodes that become most active during this propagation.

The result of a query is not only a set of documents, but also a set of relevant indexing unit (free vocabulary), people, projects and other relevant entities. All these relevant items retrieved in this manner are considered related terms that users may use to pursue (refine) their information access and knowledge navigation in a progressive fashion. For example, a retrieved "project-name" not in the original query may have strong association with the "people" named in the query.

Using the sample query and Figure 2 below, we demonstrate that there are flexible ways to formulate queries with which a user might find related items, centrally involved people, project, technology, concepts etc using the associativity of the network. The user can also input his or her feedback through a pop-up menu.

A Sample Query:

A user may issue the following query to find information about the collaborator of Enrico Motta in the area of Knowledge Modelling.

((:FREE "collaborator")(:TECHNOLOGY "Knowledge Modelling")(:PEOPLE "Enrico Motta"))

Subsequent queries (subsequent to the initial query) can be performed visually by using a pop-up menu and a navigation panel. Finding information in this way makes use of associative relations among entities and makes it possible to capture user's feedback. After the system has retrieved the network of relevant nodes, the user responds with Relevance Feedback, indicating which features are considered (by that user) relevant to the query (context) and which are not. Using a mouse, the user marks features indicating that the feature was **Relevant to Current-Context** or **Irrelevant to Current Context**. User may as well choose to provide general feedback using **Relevant in General** or **Irrelevant in General** option (c.f. Figure 2).

Not all features need be commented on by the users. However, the feedback such as **Irrelevant in General** on a statistically identified entity (which doesn't make any meaningful sense) may be useful to hide that feature permanently from the system.

This is an effective way of navigating knowledge since human knowledge is usually triggered on context and with the presentation of evidence.

The user may also navigate information using the associativity in the network by using the *Navigation Panel* (c.f. Figure 2, upper right panel).

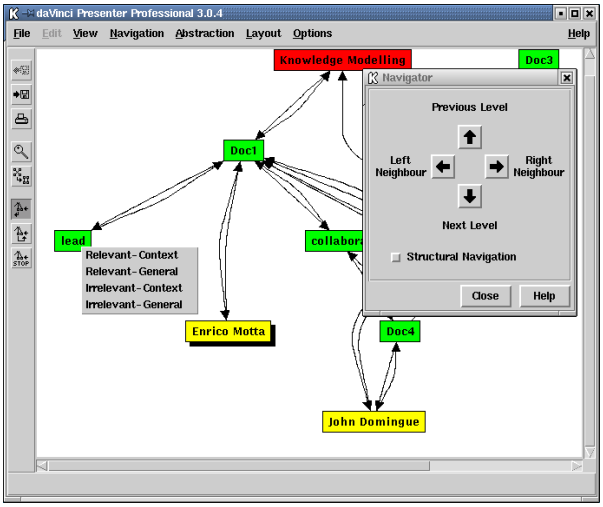

Fig. 2: Retrieval, Navigation and Feedback
Based on a user's query, a set of relevant interconnected nodes are presented to the user. The user can use the pop-up menu (left) to input feedback. The user can also use the navigation panel (right) to navigate through associated nodes.

An instance of formulation of a subsequent query directly from the feedback can be explained as follows. First, terms from the previous query are retained. Positively marked features are added to this query, while the negated features are removed. Equal weight is placed on each of these features, and the activation is propagated through the original network and the most activated nodes are retrieved and displayed.

Nodes marked by the user with positive or negative feedback act as sources of a signal that then propagates along the weighted links. A local learning rule then modifies the weight on links directly or indirectly involved in the query process.

The proposed mechanism tries to make the most direct correspondence between the connectionist notion of activity and the IR notion of relevance. The activity level of nodes at the end of the propagation phase is considered to be a prediction of the probability that this node will be judged relevant to the query presented by the user (c.f., Figure 3).

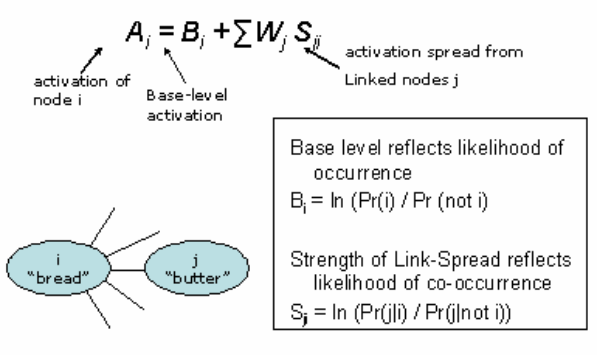

Fig. 3: Probabilistic interpretations of Spreading Activation

Weighting Scheme: Note that the initial weight of the original network is estimated from the document collection. The system's interactions with users are then considered as challenges. Given a query, the system predicts which nodes will be considered relevant and the user confirms or disconfirms this prediction. These results update the system's weights to reflect the system's updated estimates. Thus, the system produces results from the combination of two completely different sources of evidence: the related features underlying its initial indexing and the opinions of its users. *Activation retention* and *activation propagation* thresholds are important tuning parameters for a spreading activation network.

Scalability Issues: A straightforward mechanism exists to incrementally introduce new documents into the system. Links are established from the new document to all of its initial indexing units and other entities; new nodes are created as necessary. The weights on these links are distributed evenly so that they sum to a constant. Because the sum of the (outgoing) weights for all nodes is to remain constant, any associative weight to the new document must come from existing link weights.

Although easily incorporating new documents and new query terms is a valuable property for any IR system, from the perspective of machine learning these are examples of simple rote learning, and they are necessarily dependent on the specifics of the IR task domain. The main advantage creating such framework is the use of general-purpose connectionist learning techniques that, once the initial document network is constructed, are quite independent of IR tasks but can capture human expertise and knowledge within the organisations to some extent in a collaborative manner.

Current Status and Future Work

A prototype system is currently being implemented based on the above-mentioned spreading activation framework. However, only ad hoc experimentations are done with small data sets to validate only the theoretical aspects. We are yet to evaluate the system in terms of usability and performance, but this is on the top of our agenda.

We are also working on a better *Entity Recognition and Annotation Tool* since that is a crucial task to improve the performance and accuracy of this system.

At present, we are using *daVinci* APIs originally developed at University of Bremen, Germany to help us visualising the results. Future work will also include developing our own visualisation and navigation tools.

Conclusions

The spreading framework is designed to capture knowledge in the form of a trained network from three different sources of knowledge: (1) structured and inferable sources (e.g., knowledge bases, databases), (2) assertive sources (as evidenced in unstructured free text), and (3) tacit sources (as expert-feedback triggers the system to learn and adapt and system-response triggers the associative performance of the expert).

Even though the RDF and Semantic Web initiatives are already there for a few years, structured and self-descriptive data are still scarce. It is envisaged that huge amount of unstructured data will remain electronically available and serve as an invaluable source of knowledge. Any knowledge management tool therefore, should carefully consider how to incorporate knowledge available in structured, unstructured and tacit forms. Our approach is aimed at such an ambitious target.

Acknowledgements

This work is supported by the Advanced Knowledge Technology (AKT) consortium under the UK Engineering and Physical Science Research Council grant number GR/N15764/01. The authors like to thank Victoria Uren and Jiacheng Tan for their useful feedback.

References

R.K. Belew (1989). *"Adaptive Information Retrieval: Using a Connectionist Representation to retrieve and learn about documents"*, *Proceedings of 12th ACM-SIGIR Conference, pp. 11-20, Cambridge, Mass., USA.*

R.K. Belew (2000) *"Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW"*, Cambridge University Press, UK.

F. Crestani (1997). *"Application of Spreading Activation Techniques in Information Retrieval"*, Artificial Intelligence Review, 11(453-482).

F. Crestani and P.L. Lee (2000). *"Searching the Web by Constrained Spreading Activation"*, Information Processing and Management, 36(585-605).

J. Domingue, E. Motta and O. Corcho Garcia (1999) *"Knowledge Modelling in WebOnto and OCML: A User Guide"*, available from: *http://kmi.open.ac.uk/projects/webonto/user_guide.2.4.pdf*

G. Fischer and J. Ostwald (2001). *"Knowledge Management: Problems, Promises, Realities, and Challenges"*, IEEE Intelligent Systems, 16-1(60-72).

E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning, *"Domain-specific keyphrase extraction"* , Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden.

Y. Kalfoglou, J. Domingue, E. Motta, M.Vargas-Vera, and S. B. Shum (2001). *"myPlanet: an ontology-driven Web-based personalised news service"*, Proceedings of the IJCAI'01 workshop on Ontologies and Information Sharing, Seattle, USA.

A.S. Lopatenko (2001). *"Information Retrieval in Current Research Information Systems"*, Workshop on Knowledge Markup and Semantic Annotation, First International Conference on Knowledge Capture, K-CAP 2001, Victoria, Canada.

D.L. McGuinness (1998). *"Ontological Issues for Knowledge-Enhanced Search"*, Proceedings of Formal Ontology in Information Systems, Washington DC, USA.

D.L. McGuinness (2002) *"Ontologies Come of Age"*, in Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, Eds.: MIT Press, 2002.

S.E. Middleton, H. Alani, and D.C. DeRoure (2002). *"Exploiting Synergies Between Ontologies and Recommender Systems"*, Semantic Web Workshop 2002, Hawaii, USA.

E. Motta, S. Bucckingham-Shum, and J. Domingue (2000) *"Ontology-driven document enrichment: principles, tools and applications"*, International Journal of Human-Computer Studies, 52(1071-1109).

S. Preece (1981). *"A spreading activation network model for information retrieval"*, PhD thesis, CS Dept., Univ. Illinois, Urbana, IL, USA.

G. Salton and C. Buckley (1988). *"On the use of spreading activation methods in automatic information retrieval*", Technical Report 88-907, Dept. Computer Science, Cornell Univ., Ithaca, NY.

B.R. Schatz (2002) *"The Interspace: Concept Navigation across Distributed Communities"*, IEEE Computer, 35-1(54-62).

M. Vargas-Vera, J. Domingue, Y. Kalfoglou, E. Motta, and S. Buckingham-Shum (2001). *"Template-driven Information Extraction for Populating Ontologies"*, Proceedings of the IJCAI'01 workshop on Ontology Learning, Seattle, USA.

daVinci API Reference:
http://www.informatik.uni-bremen.de/daVinci/docs/reference/api/