First publ. in: Proceedings of AISB (2006), 3, pp. 98–101

Towards Associative Information Access

Michael R. Berthold*
*Konstanz University
Dept. of Computer and Information Science
Konstanz, Germany
Michael.Berthold@uni-konstanz.de

Andreas Nürnberger†
†Otto-von-Guericke Universität Magdeburg
Faculty of Computer Science
Magdeburg, Germany
nuernb@iws.cs.uni-magdeburg.de

Abstract

We propose a framework for a unifying, associative access to distributed and heterogenous information resources. The classical index generation is replaced by a process which builds associations between existing information entities and allows for an interactive exploration of information accessible through this structure. Positive ("'this looks interesting") as well as negative ("'I know this!") user feedback allows the system to quickly narrow down on interesting pieces of information. The continuous integration of new analysis engines, added sources of information and user feedback allow the formation of a corporate wide memory and expert knowledge repository.

1 Motivation

Large corporations increasingly drown in all sorts of data and other types of information they collect. Modern storage technology essentially sets no limit to the amount of information that can be stored. The huge challenge is the problem of usage — how can users be sure that they did take into account all relevant pieces of information that relate to the current task or problem they are dealing with?

One prime example for this scenario are research departments in many pharmaceutical companies. In order to successfully develop new drugs, many different types of information need to be combined, in the end resulting in a new idea for a medication that has not been patented before, that has no dangerous side effects, or that is not, in some similar form, already being explored elsewhere. Currently this process relies heavily on experts having intuition, long years of experience and hopefully the right insights at the right time. The sources of information these experts rely on are distributed across the entire company (and some also over the entire internet): experimental protocols, patent information, scientific publications, biological information about metabolic pathways just to name a few, and not to forgot, also the colleague down the hall who would have something interesting to say but who our expert did not happen to meet at the coffee pot.

Current approaches try to address this problem by building huge information repositories based on sophisticated database technology. Associative Information Networks, as described here, aim to take an alternate approach – instead of bringing all the information together we propose to build a meta structure that points to the information and helps the user find interesting associations among different pieces of information through means of exploration and context refinement. This meta structure is continuously updated as more sophisticated methods to analyze the information sources arise. In addition, it is possible to naturally incorporate user annotations, capturing expert knowledge and feedback on the way. This process is supported by methods derived from research in the areas of data mining, information retrieval, knowledge management, network and graph theory, data visualization and human computer interaction.

2 Related Work

There has been a lot of work done in the past on the idea of associative information processing, which was in the beginning mainly motivated by the associative information processing capabilities of the human brain (see, e.g., the work of Collins and Loftus (1975)). Thus we can find methods ranging from very general neural network based approaches of Kohonen (1977, 1984), over possibilistic networks or graphs (Borgelt et al., 2000; Cao, 2000) and belonging reasoning methods (Dubois et al., 1994; Gebhardt and Kruse, 1995) to very specific ideas related to document indexing and retrieval, e.g. (Chen, 1995; Chung et al., 1998; Belew, 2000). Furthermore, also ontologies and the Sematic Web (Berners-Lee et al.,

2001) might be considered as an approach to enable linking of semantically associated information.

However, several of the earlier projects failed, since almost all of them are based on the idea that it is possible to know in advance or learn automatically an almost perfect descriptive link from (index-)keywords to documents or in-between documents. This information was then used in some kind of reasoning mechanism to retrieve relevant documents. Unfortunately, in most cases this leads to the retrieval of too few or far too many documents. A further major problem had been the poor visualization methods used.

In order to circumvent these problems, more recently, some projects started in which methods have been studied that are also able to handle more general associative networks by providing interactive visualization methods. In order to navigate and browse complex association networks powerful tools for visualizing relevant subsets for the current exploration (or search) context of the user are required. Recent commercially available approaches that try to tackle this problem are, e.g., the Personal Brain (`http://www.thebrain.com/`), a navigator for indexed data that is however only able to access documents on a local data repository and the iAS KnowledgeSuite (`http://www.knowledgesuite.de/`). The KnowledgeSuite performs a semantic text analysis and creates strong links between previously identified, named entities. In this case, however, association are originate only from primed neurons using positive activity spreading. No interactive refinement or inclusion of uncertain, imprecise information is possible.

In general one might argue that the linking of documents as proposed for the semantic web might solve the problems of linking information sources. However, in the semantic web, one is forced to either link or not link documents, where an existing link has a clear, semantically valid meaning. Even though it is in general possible to introduce mechanisms for context based links (as realized for example in topic maps, see e.g. Biezunski et al. (1999)), no mechanism for storing 'gradual' (e.g. possibilistic, probabilistic, or simply anecdotal or evidential) links between documents are implemented. Furthermore, in the semantic web the whole web is seen as the knowledge base which includes both, links *and* information chunks. In our approach we add a general layer of links over (the possibly already existing link layer within) the considered database of information entities, which could consist of information in the world wide web, a local database or even notes on a local PC. This layer allows to model a personal (or group based) view on the same information, independent of (and not conflicting with) links already present in the data. However, we can easily incorporate general concepts of the (semantic) web, like URIs and existing ontologies in order to model and exploit already available information.

Another aspect that distinguishes our approach from semantic web (or more general logic based) approaches is that we do not use reasoning mechanisms that require a consistent descriptions of relations between information chunks. The main goal of the reasoning mechanism is to detect information that is most likely interesting to the user for any reasons (may be even because its contradicting!). In contrast, the reasoning mechanism itself is able to provide an explanation why some information has been proposed.

One additional differentiator is the ability for continuous learning and updating of the underlying structure. Through integration of new analysis engines, new information sources, or also manual feedback the network continuously refines it's internal structure.

3 Associative Information Networks

3.1 Structure

Associative Information Networks (AI Net in the following) consists of nodes and labelled edges. Each node represents an entity, which can be a concept from the application area (e.g. a disease, or metabolic pathway) or a named entity, such as a gene, a protein, or a specific target. Edges represent links between these entities and are labelled with a reference to the information source(s) and information about the analysis engine that created it from these sources. In addition, each edge holds a weight, modelling the strength of association, and a label indicating the type of the edge. This way, a link can potentially also be derived from an ontology, representing semantic connections between nodes.

3.2 Learning and Refinement

In order to generate the AI Net we need to introduce nodes, and links in between them. Refinement may cause adjustment of links and addition of new nodes. There are two primary ways how both, nodes and links can be added:

- automatic generation: using analysis engines, links between existing nodes can be added or modified. Each analysis engine has a particular purpose and will, for instance, find co-occurrences of words in documents, correlations of genes in gene-expression experiments, structure-activity relationships via the analysis of cell-assay images, or connections between genes and diseases from the analysis of patent information. In comparison, this would resemble the collection and modelling of automatically derivable domain knowledge. Of course, the addition of newly developed analysis engines can continuously update the network.

- manual interaction: throughout usage of the AI Net, the user is able to manually adjust weights of links, mark links as wrong, or insert new links with annotations explaining their purpose. This interactive refinement allows to capture expert knowledge and feedback on the fly and enables the system to model expertise available within a corporation. It is, of course, crucial that this interaction is handled in an intuitive way. The user should not be required to adjust numerical weights or draw links between abstract nodes.

Adding new databases, or more generally, information sources is straightforward – as long as an analysis engine is provided that produces dependencies between entities represented by nodes, new links can easily be added. One further extension of this system would also allow to generate new nodes (and node types) by analyzing external information sources.

3.3 Link Formation: Details

As described above, links can be introduced automatically or through manual refinement. The latter process can be seen as user annotations, incorporating expert knowledge into the network and are therefore mainly an issue of user interface. In the following, we briefly outline, based on a number of examples, how the automatic generation of links and link-weights works.

- semantic links: these are strong links (usually weight $= 1.0$) which are derived from well-known structures, such as ontologies or semantic networks. Those are usually created by an expert. Semantic nets, as extracted (semi-) automatically from data will need to add a component that computes the confidence for each link and convert this to a weight.

- syntactic links: these are links that are generated by a shallow analysis of data. The most prominent example would be a text parser that converts words to stems, eliminates fill words and then produces a set of bi- or trigrams. The corresponding nodes in the AI Net will be connected by weak links. For an example of the corresponding weight computation, see below.

- anecdotal evidence: These are links set by a user, creating links for hypotheses generated by a user (or based on hear-say). Weights of such links are generally low. These links are in contrast to expert-based annotations that generally have very high weights.

- data driven links: These types of links will constitute the vast majority of network weights in most cases. They are generated automatically from data repositories. A few example (here for the context of a pharmaceutical AI Net) could be:

  Gene correlations derived from gene expression data. Links are introduced when, for example, a specific threshold θ for co-occurrence in experimental data is surpassed. The link's weight reflects the correlation strength and for more than two-dimensional correlations the corresponding multi-edges are introduced. In addition each of these links will carry an annotation pointing to the source of it's weight, in this example a link to the experiment and some meta information (threshold θ, date of analysis, reference to exact computation of weight).

  Textual analysis where co-occurrence of named entities within a specific distance (= words in between) results in a weak link to be introduced. The weight depends on distance and quality of text source.

  Links between gene and protein names derived from scientific articles based on a bigram analysis. Weights are derived from the average distance and frequency of occurrence in documents, analogous to the TFIDF-score (Term frequency / inverse document frequency).

- ontology/thesaurus links: Based on an existing ontology links will be introduced to connect entities that are related based on this ontology. This resembles a 1:1 correspondence between each link in the ontology and a link in the network. The resulting links are strong links, i.e. carry a weight of 1.0 since there is (usually) no doubt about the reliability of that particular

piece of knowledge. Otherwise it would need to be reflected in the link's weight.

Obviously many other types of links can be generated, since the underlying structure is invariant to origin or meaning of links.

3.4 Exploration: Finding interesting associations

The network's structure can be used in various ways to find potentially interesting pieces of information. Most straightforward would be the search for tightly connected other entities, such as another gene that is related to the ones the user just saw within an experiment. This can be implemented via a simple neighbor-search in the network, finding all genes that are connected to the set of "query" genes.

More powerful are, however, searches that find related pieces of information via various steps, or so-called bridge concepts. This can be implemented analogous to activity spreading methods, as known from the neural network community (Cohen and Kjeldsen, 1987). The real power, in the concept presented here, lies in the ability to perform this search interactively. Throughout the search the user can weight entities that he finds interesting positively (and the ones he does not care about negatively), instantly affecting the activation pattern and hence the associations the network proposes. Such an interactive scheme will heavily rely on a suitable visualization of the graph network (see, e.g. Chen (2004)) and appropriate adaptive user interfaces.

4 Conclusions

In this paper we have briefly presented the idea of a generalized associative information network. With this concept we try to simulate aspects of the associative capabilities of the human brain in order to support a user in gathering information about a specific problem at hand. The tool is not meant to offer problem solving capabilities, but rather to point out information pieces a user might have otherwise not had the chance to look at, be it for lack of knowledge about their existence or because of a failure to see their importance for the task at hand.

References

R.K. Belew. *Finding out About*. Cambridge University Press, 2000.

T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May, 2001.

M. Biezunski, M. Bryan, and S.R. Newcomb, editors. *ISO/IEC 13250:2000 Topic Maps: Information Technology – Document Description and Markup Languages*. ISO/IEC, 1999.

C. Borgelt, J. Gebhardt, and Ru. Kruse. Possibilistic graphical models. In G.D. Ricci, R. Kruse, and H.-J. Lenz, editors, *Computational Intelligence in Data Mining*, pages 51–68. Springer-Verlag, Wien, 2000.

T. H. Cao. *Fuzzy Conceptual Graphs: A Language for Computational Intelligence Approaching Human Expression and Reasoning*, pages 115–120. Physica-Verlag, Heidelberg, 2000.

C. Chen. *Information Visualization*. Springer, 2004.

H. Chen. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. *J. Am. Soc. Inf. Sci.*, 46(3):194–216, 1995.

Yi-Ming Chung, William M. Pottenger, and Bruce R. Schatz. Automatic subject indexing using an associative neural network. In *DL '98: Proceedings of the third ACM conference on Digital libraries*, pages 59–68, New York, NY, USA, 1998. ACM Press.

P.R. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management*, 23(2):255–268, 1987.

A.M. Collins and E.F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.

Didier Dubois, Jerome Lang, and Henri Prade. Automated reasoning using possibilistic logic: Semantics, belief revision and variable certainty weights. *IEEE Trans. Data and Knowledge Engineering*, 6 (1):64–71, 1994.

Jörg Gebhardt and Rudolf Kruse. *Reasoning and Learning in Probabilistic and Possibilistic Networks: An Overview*, volume 912 of *Lecture Notes in Artificial Intelligence*, pages 3–16. Springer-Verlag, Berlin, 1995.

Teuvo Kohonen. *Associative Memory - A System Theoretic Approach*. Springer-Verlag, Berlin, 1977.

Teuvo Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 1984.