

h-Techsight: a Knowledge Management Platform for Technology Intensive Industries

A. Kokossis^{a*}, R. Bañares-Alcántara^b, L. Jiménez^c and P. Linke^a

^aDepartment of Chemical and Process Engineering, University of Surrey
Guildford, Surrey GU2 7XH, UK

^bDepartment of Engineering Science, Oxford University
Parks Roads, Oxford OX1 3PJ, UK

^cDepartment of Chemical Engineering and Metallurgy, University of Barcelona
Martí i Franquès 1, Barcelona 08028, Spain

Abstract

In knowledge-intensive industries it is of crucial importance to keep an up-to-date knowledge map of their domain in order to take the most appropriate strategic decisions. The main objective of the knowledge management platform (KMP) is to improve the capabilities of chemical process industries to monitor, predict and respond to technological trends and changes. The search, retrieval, analysis, filtering, rating and presentation of information retrieved from the web (or any other type of document) are elucidated through the use of multi-agent systems, dynamic ontologies and learning techniques (conceptually similar documents are clustered and natural language processing techniques are used to retrieve new terms). Discovery of new knowledge leads to recommendations of modifications in the ontology (either classes or instances) by pruning irrelevant sections, refining its granularity and/or testing its consistency. The KMP works using an intelligent, asynchronous and concurrent process to achieve high quality results.

Keywords: knowledge management; knowledge retrieval; web search; ontology.

1. Introduction

Decision making in technology intensive industries has to be made based on information that is constantly evolving. New technologies, markets and products emerge and change, and relevant information can be found only if one knows exactly where to look for it. Unfortunately, the amount of information and the various ways in which it can be presented, makes the retrieval of useful information an increasing more difficult and work intensive task.

A KMP to monitor, predict and respond to technological, product and market trends has been developed (h-Techsight, 2001; Stollberg et al., 2001), which innovative points are:

* Author/s to whom correspondence should be addressed: A.Kokossis@surrey.ac.uk

- *h-TechSight* performs the search based on an initial ontology supplied by the user. An ontology is a conceptualisation of a domain. Ontology-based search is more accurate and complete than traditional keyword-based search (Fensel, 2001).
- *h-TechSight* has the capability to suggest desirable modifications to the initial ontology based on the information retrieved (in the web or the databases). We refer to this capability as *dynamic ontologies* because it provides a mechanism to update the understanding of a domain with the available, ever-evolving, information.

h-TechSight KMP can operate in two different modes: as a generic search or as an application search tool. In the generic search mode the system uses the whole web (or a selected domain in the web) as an information source. Search is performed by a multi-agent system and the links retrieved are analysed using text analysis techniques and clustered into new categories. In the application search mode the system searches in domains where the information, while unstructured can be found in documents of similar patterns. The smaller number of records and their similar format permit the application of powerful analysis tools (GATE and WebQL).

2. Generic Search Mode

The generic search mode architecture (Figure 1) is based in four different modules: the ontology editor, the multi-agent search system, the clustering search system and the dynamic ontology update.

2.1. Ontology editor

Under the generic search mode, a ontology editor has been integrated in the KMP to facilitate the creation, customisation, browsing and modification of ontologies.

Each user of the KMP has a personalised area in which his/her ontologies are stored, thus versions of the same ontology are stored to further analyse their dynamics. Uploading and downloading of ontologies are always performed in RDF format.

2.2. Multi-agent search system

This module receives as an input the ontology and uses search engines to perform semantic based search, according to a predefined set of searching parameters. In this way, the Multi Agent Search System (MASH) finds web pages that contain relevant information to each concept in the domain of interest (described by the class path, as each class inherits all instances defined in their ancestors). The retrieval, rating and filtering processes are performed asynchronously, concurrently and in a distributed fashion by the agents with different roles. MASH is described in detail elsewhere (Banares-Alcántara et al., 2005).

2.3. Clustering search module

It is used to perform analysis on the results received from the MASH to propose new categories. For each URL provided by the MASH system, this module finds the URLs that point to it. Let A, B and C be three incoming links of URL D (Figure 2). The module extracts keywords from the incoming links, processes their contents and extracts terms from their hyperlinks to D. Each set of terms that corresponds to a URL D is mapped to a set of concepts of the ontology. *WordNet* (Wordnet, 2004), an online lexical reference system, in which English nouns, verbs, adjectives and adverbs are

organised into synonym sets, is used for this purpose, and thus, the system is able to process html, pdf, doc or xls documents. The procedure is as follows:

- For each term (t_i) in the set, a clustering mechanism finds the closest ontology concept (c_j) in *WordNet*.
- Extracted terms are mapped to *WordNet* (t_1 is mapped to nodes $t_{1,1}$, $t_{1,2}$ and $t_{1,3}$).
- Ontology concepts are mapped to *WordNet* (c_1 is mapped to nodes $c_{1,1}$ and $c_{1,2}$).
- The distance between every node of t_1 and c_1 is computed using the Wu and Palmer distance (1994).
- The closest pair of nodes ($t_{1,x}, c_{1,y}$) defines the distance between term t_1 and concept c_1 .

After this process, each URL is described by a set of terms and a set of ontology concepts. Clustering is performed using the set of concepts of each URL using the DBSCAN (Density Based Spatial Clustering of Applications with Noise) and a similarity measure between URLs (Ester et al, 1996).

- For each document D of a cluster the neighbourhood of D has to contain at least a minimum number of documents (MinDoc).

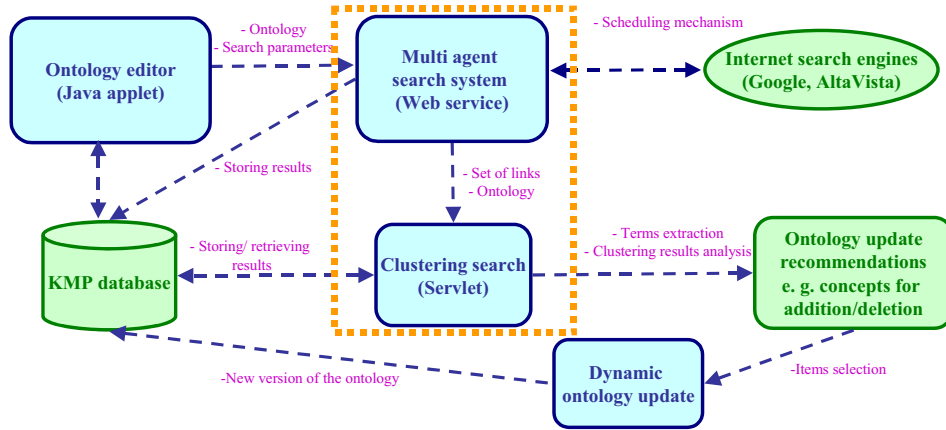


Figure 1. Generic search mode architecture.

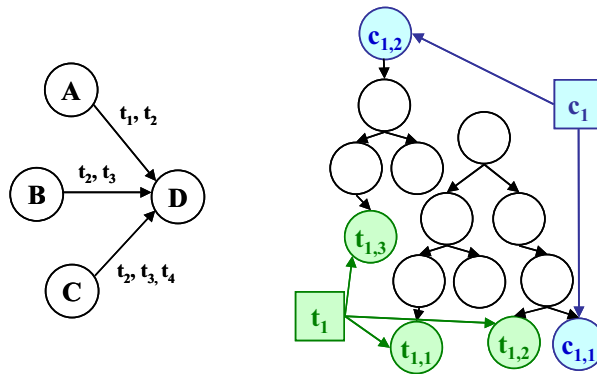


Figure 2. (a) URL relationships; (b) where A , B , C and D are URLs, t_i and $t_{i,j}$ are terms and c_i and $c_{i,j}$ are concepts.

- The neighbours of D are defined to be all the documents whose similarity to D is higher than, or equal to, a minimum similarity threshold (*MinSim*).
- DBSCAN is able to detect clusters with strange geometries. The number of clusters is not predefined.
- The algorithm is repeated for different values of *MinDoc* and *MinSim* and the scheme that provides the most compact and discrete clusters is selected.

Dynamic ontology update

Based on the clustering results, the user has the ability to extend/modify the ontology with the newly discovered keywords. In both search modes the system presents a table to the user where she/he can choose which items should be added to the ontology, and if the recommendation is to add the new terms as new classes or as instances of existing classes.

Each time the user saves an ontology, a new version is created and stored in the database. Therefore, a user is able to return to a previous version of the ontology at any point in time. The versioning mechanism allows the user to keep track of the modifications applied to an ontology (as stated in section 2.1).

2.5. Scheduler mechanism

The scheduling sub-system of the KMP gives the ability to the users to schedule both search modes of the platform and view the results at a later stage. The scheduled searches are fully configurable (e. g. search parameters of the MASH/Toolbox). This facility opens up the possibility to apply the KMP functionalities not only to snapshots of the web but as it evolves in time.

3. Application Search Mode

In the application model the user defines a set of URL sites or database of documents that contain relevant information. The user may wish to automatically extract information, classify documents according to the ontology, assess the “dynamics” of the domain (e.g. site), and monitor changes. The application search mode is illustrated in Figure 3. A number of dedicated Natural Language Processing (NLP) tools are employed.

3.1. GATE

GATE is an architecture for NLP, that enables an automatic semantic annotation of web mined documents using the terms of the ontology (entities or relationships). The KMP uses GATE to support the evolution of instances. The findings are stored to perform statistical analysis in order to monitor trends over time. GATE is described in detail elsewhere (Maynard et al., 2004).

3.2. Toolbox

This module applies NLP techniques to assess and validate ontologies. The latter are usually developed in an *ad-hoc* fashion or are recycled without a prior assessment for relevance and quality. Modules in the toolbox can be applied to enhance ontology relationships, discover new terms, and adjust ontology components according to the application context.

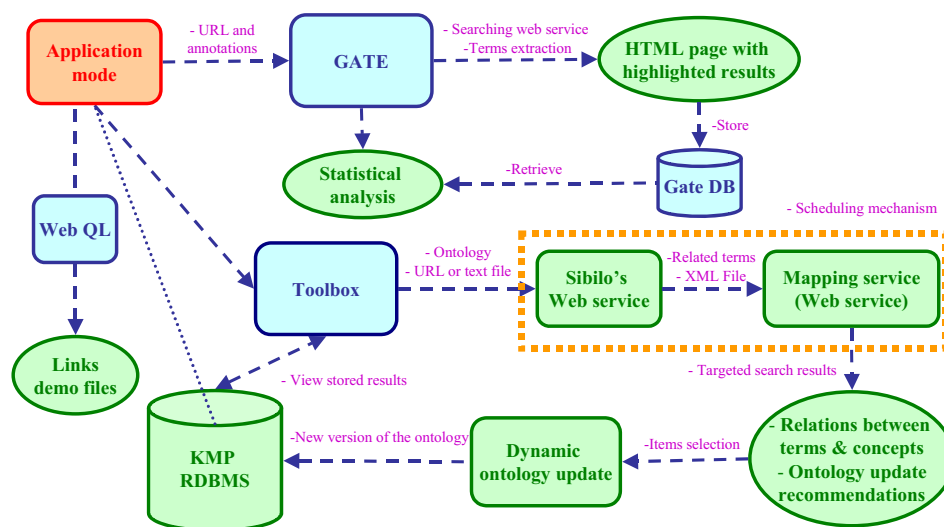


Figure 3. Application search mode description.

3.3. WebQL

WebQL (WebQL, 2005) is a commercial web mining tool used to collect unstructured data and prepare them for analysis. WebQL has been selected out of several packages.

3.4. Contextual ontologies

Contextual ontologies are developed “*on-the-fly*” out of an ontology superset and a knowledge domain of preference. Contextual ontologies are descendants of the parent ontologies and combine concepts, relationships and properties of each parent. A contextual ontology in “*Patenting catalysts*” can be thus developed out of a “*Chemical Engineering*” ontology, a “*Materials*” ontology, and a “*Patent Office*” ontology.

4. Knowledge Management Platforms Developed

A number of case studies have been developed in diverse areas (e.g. technologies, employment, health services), as shown in Kokossis and Bañares-Alcantara (2003). These include process engineering applications (mostly with h-TechSight partners), chemical engineering employment applications (in collaboration with IChemE), and biomedical applications (in collaboration with the Parkinson’s Disease Society).

4.1. Employment markets

The KMP prototype was used to create a portal for employment markets, as this tool is useful to help individuals or organisations to monitor trends in demands for a particular areas of expertise to enable a systematic continuing professional development. The portal (<http://prise-serv.cpe.surrey.ac.uk/hTechSight/>) aims to support engineers with a wish to understand the job market and search for a job, companies and SMEs to subcontract/undertake projects, academic institutions preparing students for the job market, and professional institutions. The dynamic change identified in employment markets is continuously incorporated in revised ontology versions that reflect up-to-date knowledge in the field.

4.2. Parkinson portal

The portal makes use of an extensive knowledge model for Parkinson's Disease. The model is an extensive ontology modelled in DAML+OIL and developed at the University of Surrey with the collaboration of medical doctors, practitioners, communities and local hospitals. The portal site can be found in <http://prise-serv.cpe.surrey.ac.uk:8080/>.

5. Conclusions

This paper describes the results obtained with the prototype of the KMP. The main objective of the platform is to improve the capabilities of chemical process industries to monitor, predict and respond to technological, product and market trends and changes. The analysis, modification and presentation of information retrieved from the web (or any other type of resource) are achieved through the use of multi-agent systems, data mining, natural language processing and clustering analysis. Ontologies are used to specify and represent the knowledge, and the KMP proposes a set of terms to be introduced as new classes or as instances of existing classes.

Preliminary user evaluation and feedback (ICHEME and Bayer AG) are positive, but more improvements are needed to improve the usability of some aspects, to clarify terminology and to remove false positives.

References

- Bañares-Alcántara, R., L. Jiménez and A. Aldea, 2005, Multi-Agent Systems for Ontology-Based Information Retrieval, European Symposium on Computer Aided Process Engineering , Barcelona, Spain.
- Ester, M., H.-P. Kriegel, J. Sander and X., 1996, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).
- Fensel, D., 2001, Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Heidelberg, Germany.
- h-TechSight, 2001, IST Project (IST-2001-33174). Available at <http://prise-serv.cpe.surrey.ac.uk/techsight> (Access: January 2005).
- Kokossis, A. and R. Bañares-Alcantara, 2003, Dynamic Information Management for Web-enabled Environments in the Chemical Process Industries, FOCAPO, USA.
- Maynard, D., M. Yankova, N. Aswani and H. Cunningham, 2004, Automatic Creation and Monitoring of Semantic Metadata in a Dynamic Knowledge Portal, 11th Int. Conf. on AI Methodology, Systems, Applications-Semantic Web Challenges.
- Stollberg, M., A. Zhdanova and D. Fensel, 2001, H-TechSight: a Next Generation Knowledge Management Platform, J. Inf. and Knowledge Manag., 3 (1): 47-66, 2004.
- WebQL, 2005, Available at <http://www.ig.com.au/WebQL.htm> (accessed January 2005).
- Wu, Z. and M. Palmer, 1994, Verb Semantics and Lexical Selection. 32nd Annual Meeting of the Associations for Computational Linguistics, 133-138, Las Cruces, New Mexico. <http://citeseer.ist.psu.edu/wu94verb.html> (accessed January 2005).
- WordNet, A Lexical Database for the English Language. Available at <http://www.cogsci.princeton.edu/~wn/> (accessed January 2005).

Acknowledgements

h-TechSight is a Framework V EU funded project (IST-2001-33174).