Market Blended Insight: modeling propensity to buy with the Semantic Web

Manuel Salvadores, Landong Zuo, SM Hazzaz Imtiaz, John Darlington, Nicholas Gibbins, Nigel R Shadbolt, and James Dobree

Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
{ms8,lz,hsmi,jd,nmg,nrs}@ecs.soton.ac.uk
ProspectSpace Ltd
25 Landsdowne Gardens London, SW8 2EQ, UK
james@prospectspace.com

Abstract. Market Blended Insight (MBI) is a project with a clear objective of making a significant performance improvement in UK business to business (B2B) marketing activities in the 5-7 year timeframe. The web has created a rapid expansion of content that can be harnessed by recent advances in Semantic Web technologies and applied to both Media industry provision and company utilization of exploitable business data and content. The project plans to aggregate a broad range of business information, providing unparalleled insight into UK business activity and develop rich semantic search and navigation tools to allow any business to 'place their sales proposition in front of a prospective buyer' confident of the fact that the recipient has a propensity to buy.

1 Introduction

The Market Blended Insight project (DTI Project No: TP/5/DAT/6/I/H0410D) is a three year applied research project funded under the UK Governments Technology Programme.

The innovation challenge for the project is: to overcome the problem that traditional marketing techniques have broad push without knowing if the recipient has a propensity to buy. The project is extending world class Semantic Web research from the EPSRC's "Advanced Knowledge Technologies IRC" [1] and applying it to a large scale collection of real life UK company data sources to understand if an organization's propensity to buy can be discovered within a very large pool of information.

To ensure the research is undertaken in a real world scenario the project has direct involvement from marketing departments within UK businesses. The consortium for this project includes marketing departments of the following companies: ParcelForce Worldwide, British Gas Business, AXA, Clydesdale and Yorkshire Bank (NAGE) and 3M. Based on their marketing needs the project is developing advanced methods of analysing target markets and the innovation includes:

- the anticipated scale of the information source we plan to create, based on the 3.7 million companies that constitute the entire UK economy.
- The complexity of the collection of ontologies, covering a rich depth of information for each company.
- Finally and most difficult, the innovation required to identify within the information the semantic relations and queries required to determine propensity to buy given a sales proposition.

The project is building an industrial scale prototype that aims to provide UK business users with dynamic, relevant insight into their target markets and allows them to make informed decisions on a potential clients propensity to buy.

The first phase performed a needs analysis within the consortium which developed the following classification of marketing processes: strategy, scoping, scanning, data processing and interpreting. Each consortium member has described the different areas and their needs for each in terms of desired outcomes.

A subset of the expected outcomes has been achieved by deploying a number of scenarios in a prototype architecture. This paper discussed two scenarios from the first prototype: *Micro Segmentation* and *Value Chain*. The implementation of both scenarios has tackled technical issues related to Semantic Web such as: semantic annotated data extraction, ontology driven data integration, data generation through onotology reasoning and ontology driven data visualization.

Besides these scenarios and the architecture the following sections also describes related work, the information extraction process and conclusions.

2 Related Work

There is an existing market for Business Intelligence tools and the desire for integration of information services on demand and their descriptive metadata to further improve the performance of corporations in the B2B market is described in [32]. Researchers in the Semantic Web community have been working on exploiting semantic technology for the integration of public information in a wide range of application domains with knowledge held in heterogeneous formats, representations and structures [14]. The research into Semantic Web technologies is diverse, developing emergent middleware frameworks in areas such as service composition [27, 35, 26], data integration [29, 1, 31] and data extraction. In following sections it is shown how this project has made use of several data extraction techniques for harvesting the required data and how semantic data integration is one of the basic pilars to build up the scenarios.

The mashing-up of knowledge was previously demonstrated by the AKTivePSI project [2]. The Semantic Web technologies were widely used to integrate both data and metadata of company information with other knowledge from different domains, (for example, spatial knowledge from Ordnance Survey and administrative information from local councils) creating an enhanced view of the market and more efficient access of information from different perspectives [20]. The aggregation of information from structured private data sources such as RDBMs

in corporate environments with publicly available sources on the web offers an ability to create an aggregated vision of B2B market analysis where information on the same market is established from a variety of sources.

There has been progress in the extraction of data from public domains like the Internet. SPHINX[30] can be used to develop customized web crawlers. Typical crawling based on keywords has been improved with ontology driven aproaches [25]. Other solutions allow for the extraction of tree data structures and the semantic annotation of the extracted data, manipulation of the tree structures, recording any changes and watching for any modifications [28]. Armadillio [22] uses an alternative strategy dealing with web extraction in a largely automated way that does not require manual annotations. It utilizes information that has been extracted from highly reliable source to train the information extraction.

GATE[24] is an open source architecture that provides an infrastructure to develop software components to process natural language and a GUI development environment. It contains built in components to perform tasks from simple tokenisation to complex semantic tagging and includes a Java Annotation Patterns Engine(JAPE) [23], an annotation comparison tool with performance metrics and plug-in for external parsers etc. Extraction of semantic relations from natural language text can be completed via shallow Jape patterns [34] combined with dependency parsers to identify linguistic triples that can then be mapped to RDF statements using ontology concepts and relations. The entities in the text can be identified by their instantiation in the ontology and the relation between them.

3 Application architecture

The project architecture is structured with four different components: the web user interface, the Social Network Services (SNS), the Semantic Query and Inference Services (SQIS) and the Common Extraction and Translation Framework (CETF) (see figure 1). The SNS is planned to be built in further prototypes; SQIS , CETF and the Web Portal components are detailed in this section.

The end user application is a web portal developed using Java Server Faces (JSF) [6] which renders in the user's browser XHTML [19]. UI widgets such as trees, tables, lists, maps and network have been developed extending JSF. This set of widgets are able to display the RDF/OWL [11, 17] data pulled out from the SQIS. The data is displayed for each scenario by gathering several widgets into one or more web pages. Since it is expected to add more use cases in next iterations, re-utilization of UI compononents gives a flexible and quick method for setting up new data visualizations.

The SQIS component makes wide use of Semantic Web technology to provide the ontology inferencing and data query services to the application. Both reasoning and query functionality are built on top of Jena framework [21] but it has been developed to isolate the application from underlying services in other that other frameworks can be investigated in future prototypes.

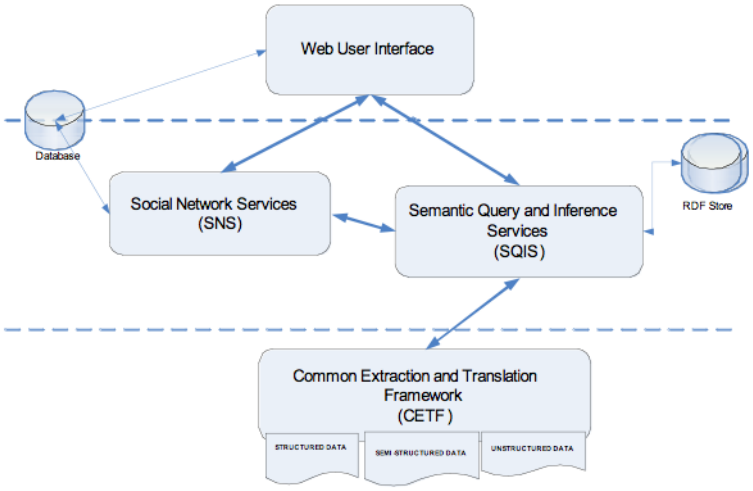

Fig. 1. General architecture diagram

The scenarios deployed in the application make use of the SQIS by defining the queries, in SPARQL [13], and the inference rules. Inference within the SQIS works in two different modes: online and offline. Depending on the specific use case need, inference that requires reasoning over massive amounts of data is performed in offline mode otherwise it is online for results requiring immediate user interaction. Both online and offline reasoning are required for Micro Segmentation and Value Chain scenarios in order to integrate different data sources as well as to prepared the data to be queried by end user.

The CETF component extracts data and translates it into RDF format. The data sources listed for each of the scenarios are extracted by this component. Currently we classify different data sources in three groups: structured, semi-structured and unstructured data, as it shown in section 4. Each time the CETF extracts a new data source, it notifies the SQIS and the new data source is registered. From that moment the data is avalaible to be used by the inference and query functions implemented in the SQIS. The CETF is built on top of a plugin architecture, see figure 2. Different plugins can be attached to the CETF in order to pull out different information, currently it supports plugins for GATE [24] and UIMA [4]. The CETF also provides crawlers for exploring and scheduling for planning iterative extractions.

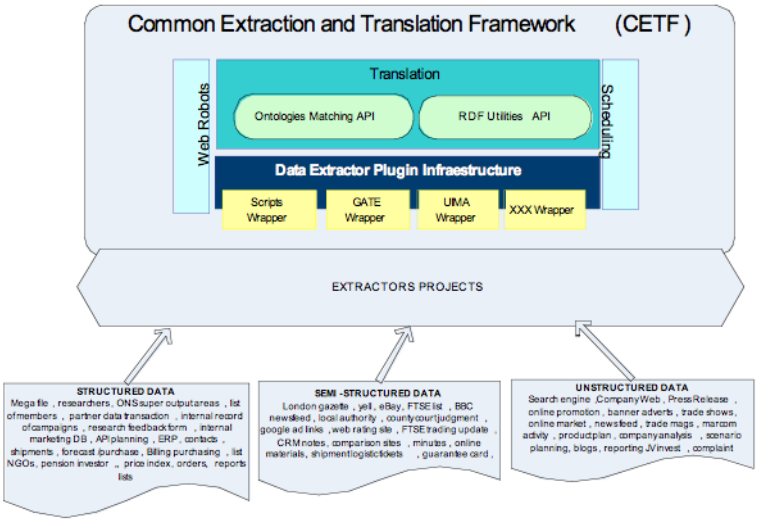

Fig. 2. CETF component

4 The information extraction process

To date this project has focused on text based information extraction. It considers "structured data" to refer to the type of data stored in databases with associated metadata reflective of the data schema. Unstructured data does not have a defined structure or schema associated with it for example free text within web pages. Semi-structured is defined as portions of the data have associated structure and meta-data or schema and portions have no meta-data. In our definition the form of structure is not considered i.e. in a HTML web page formatting instructions are helpful for processing the contents of a page but it does not contain the necessary semantics.

The amount of relevant unstructured business data is growing, and will continue to grow in the foreseeable future. According to TDWI [16] estimations of the unstructured and semi-structured data are 53 percent of an organizations overall data. Aware of this opportunity some of the organizations are increasingly feeding unstructured data into their data warehousing and business intelligence processes such as wikis, RSS feeds, instant messaging transcripts, and document management systems, e-mails, office-suite documents, Web pages, and Web logs (or blogs). The amount of e-mail data fed into these processes grew by nearly

one-half (47 percent) over the last three years, followed by office-suite documents (35 percent), Web pages (35 percent), and Web logs (27 percent) [16].

The project uses the WebSPHINX [18] web crawler for data extraction which is a implementation of the SPHINX interface [30]. The project has extended the crawler to be switched to a more topic-based focused crawling mode where the crawl is controlled by rating the links based on a high level background knowledge such as an ontology [25], and to direct the crawl for visiting specific links and data patterns. Here the crawl is started by specifying one or several start URLs and an initial ontology.

Once these parameters are specified, the extraction rules extract from the page specific HTML markup any interesting pieces of textual information. In case of PDF documents, this consists of extracting XML markups which comes as an output of the Pdfbox tool [8].

Once the desired pieces of text are extracted, the documents are processed using a GATE [24] pipeline. A GATE pipeline is application dependent but usually consists of stages of tokenisation, gazetteer lookups, pattern matching by JAPE grammars, part of speech tagging and dependency parsing etc. The project has created gazetteers of company names and street names from the company backbone which are used for detecting these entities within the documents. The Jape scripts are selected based on needs of the current extraction and identify items from post codes, phone numbers, emails, URLs etc. to the council planning application numbers and dates.

5 Use case : micro segmentation

The UK standard industrial classification of economic activities SIC(92) [15] has been widely use in marketing analysis due to the fact that its stated objective is to provide the UK with a uniform framework to classify business establishments by the type of economic activity in which they are engaged and also because the information is gathered and published by government.

One form of marketing analysis is to segment, or classify, potential customers by clustering those with common needs. A course segmentation of needs is often defined by creating clusters of organizations with the same business activity. Consortium members have agreed that SIC(92) is not a fine enough classification of business activity, under representing the activity of an organization. For instance it is not possible to find out whether a restaurant is an Italian restaurant or not. The finest SIC(92) classification is “Unlicensed restaurants and cafes”, “Take away food shops” and “Licensed restaurants”. Many organizations declare more specific details about the activity in which they are engaged in there communication to the market in order to attract the right type of clients. For instance, finding out the type of a restaurant is easily answered by searching the Web or looking at a directory service such as Yellow Pages.

The aim of this scenario is to provide detailed classification, or micro-segmentation, by extending SIC(92) classes with sub-classes that are defined by external data sources using Semantic Web techniques.

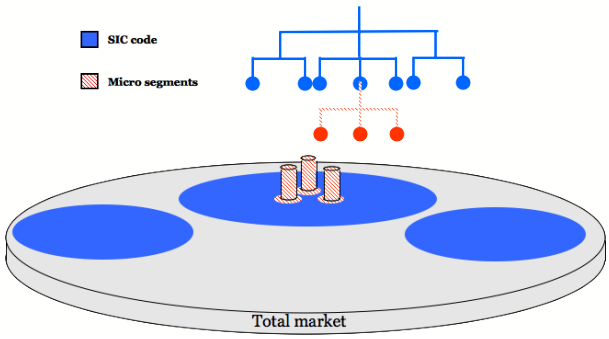

Fig. 3. Micro segmentation process

The data sources involved in this scenario are:

- A backbone of the UK companies within the London boroughs of Lewisham and Camden [9]. This data backbone collects information of more than 83,500 companies in more than 12 million RDF triples.
- SIC(92) hierarchy classification [15], a hierarchy of more than 6k nodes, this information is stored in 61,704 RDF triples.
- Ordnance Survey Point of Interest database (PointX) [7]. PointX dataset contains around 3.9 million geographic and commercial features across Great Britain. The data used in this scenario is the collection related just to the London boroughs of Lewisham and Camden.
- `www.mycamden.co.uk` [5] is a website provided by the London borough of Camden which publishes a directory, a form of classification, for companies in that specific area. The data extracted from MyCamden contains information of 1,824 companies gathered in 85 different classes. This information is stored in 93,989 RDF triples and has been extracted with the information extraction techniques detailed in section 4.

The micro segmentation use case mainly relies on two processes: 1) creating a matching between the companies listed in the additional data sets and their equivalent entity in the backbone listing of UK companies and 2) the extension of the SIC(92) classification by attaching additional sub-classes from the classification structures contained in the additional data sets. These classifications are finer than SIC(92), for instance Mycamden data provides relevant information on whether a restaurant is a Chinese or an Italian restaurant.

Both PointX and Mycamden provide additional data about companies. These data sets need to be attached to the company backbone. The company matching process creates a semantic link between every company extracted from either PointX or Mycamden to the company backbone [9]. In terms of the Semantic Web this link between the backbone and the external data sources is created through a `sameAs` OWL [17] arc. The inference to see whether two companies in different data sources are the same is achieved by comparing their names

and addresses. The rule grammar embeded within thre SQIS for this case is as follows:

```
(?compA rdf:type <mbi:company>) (?compB rdf:type <mbi:company>)
(?compA mbi:hasName ?nameA) (?compB mbi:hasName ?nameB)
equal(?nameA, ?nameB)
(?compA mbi:hasAddress ?addrA) (?compB mbi:hasAddress ?addrB)
equalAddreses(?addrA, ?addrB)
->
 ?compA owl:sameAs ?compB
```

`equalAddreses` is a built-in primitive developed within Jena to determine with a defined level of confidence if two addresses with different formats are the same.

The finest SIC(92) segments within the restaurant business activity are "unlicensed restaurants and cafes", "Take away food shops" and "licensed restaurants". As it is shown in figure 4, Mycamden and PointX provide finer classification with micro segments. In this example, restaurant business activity has been extended with 40 micro segments (30 from PointX and 10 from MyCamden).

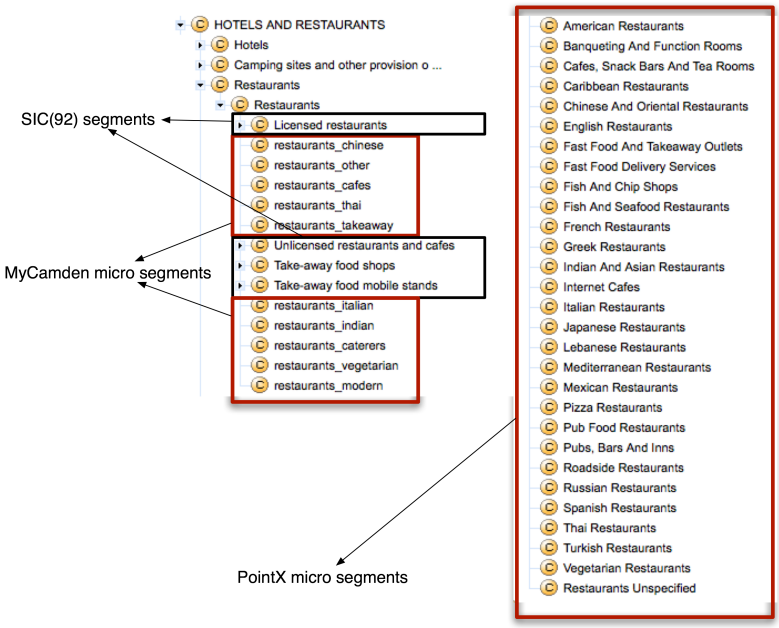

Fig. 4. SIC(92) extension for restaurants business activity

SIC(92) is stored in the system in RDF using SKOS [12]. SKOS is a Semantic Web standard that extends RDF with an specific vocabulary for knowledge or-

ganization systems and classification schemes. Using SKOS, SIC(92) is extended sub-classes from additional data sets by adding `narrower` arcs.

In our work to date, new information for 5,014 companies (4,406 from PointX and 608 from Mycamden) has been semantically integrated with the companies data backbone, providing 843 micro segments (777 from PointX and 66 from Mycamden). Once the process of creating new micro segments via extending the SIC(92) classification scheme is completed by one user of the system all end users can browse within the web application via the extended hierarchy of micro segments which have finer information about the business activity for the 5,014 companies, see figure 5.

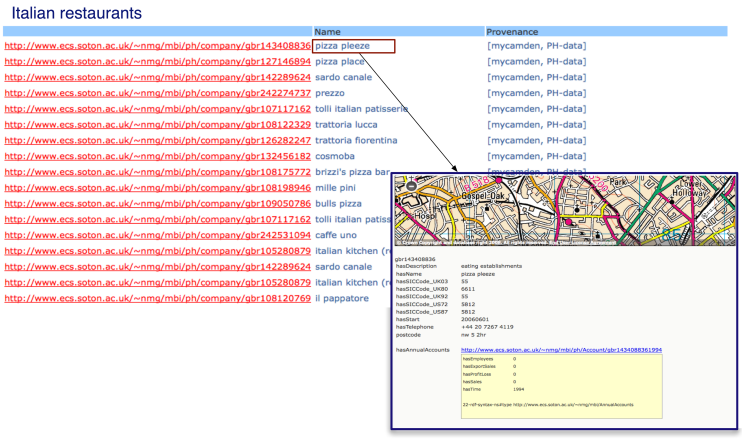

Fig. 5. Micro segmentation query interface

6 Use case: value chain

A Value-Chain is defined by Porter [33] as a series of value-generating activities. Products pass through all activities of the chain in order, and at each activity the product gains some value. The value chain framework quickly made its way to the forefront of management thought as a powerful analysis tool for strategic planning. Clear visualization of value-chain information is one of most desired scenarios from MBI consortium users. They require a tool that allows easy visualisation and manipulation of relationships between companies in order to evaluate their propensity to buy.

he consortium members decided that the initial focus for the prototype would be the relationships within the Building and Construction (B&C) industry. Domain specific data was obtained and a generic solution developed to support ontology-driven data extraction and user-view visualisation based on Semantic

Web technology. The value-chain use case is composed of four main functional components, 1) Information Harvesting, 2) Semantic Integration, 3) View Adaptation and 4) Network Visualization (see figure 8).

The Architects Journal [3] is a major web portal providing rich information on the building and Construction Industry including projects, companies and products and all their relations. The prototype harvested information from the web pages and recovered the relating transactions details for 4000 suppliers, 6000 products and 600 projects and all transaction relations using information extraction techniques detailed in section 4.

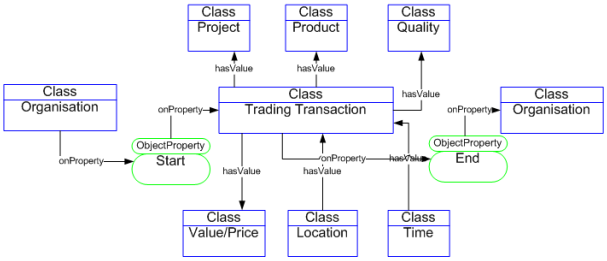

Fig. 6. Pre-inference raw data view

To harvest information the system tries to extract all the specific transaction details in order to form a solid company and project backbone about the construction industry. The data is integrated with relevant information from other sources such as those containing representation of product taxonomy and business activity hierarchy. The integration allows more restricted relations across different domains in order to support filtering out any data that is not of interest in analyzing model of value chain on the users perspective. The network visualisation of value chain is controlled by the user. The user preference is stored in a user ontology and rule syntax that supports the reasoning process necessary to generate a view of value chain model. The generic rule-based inference engine of Jena is deployed to support the knowledge adaptation of raw data model on user demand. The following example shows how a simple user view (see figure 7) is generated from raw data model (see figure 6).

The figure 6 shows a generic schema containing transaction details between any pairs of companies. The relations between companies are classified into types such as, buy, sell, service, offering, shipment, partnership, etc. The conceptual model in figure 7 has addressed a specific user interesting about "Client", "Architect", "Contractor", and "Supplier" in B&C industry, where those concepts and relations may not exist in the raw data model like figure 6. The system has created the following rule syntax to adapt the gap.

```
[Client:  (?n mbi-vc:launch ?t) (?t rdf:type mbi-vc:Investment)
(?n rdf:type aj:Organization)
```

```
-> (?n rdf:type mbi-vc:Client) ]
[Supplier: (?n mbi-vc:launch ?t) (?t rdf:type mbi-vc:Supply)
(?n rdf:type mbi-vc:Architect )
 -> (?n rdf:type mbi-vc:Supplier) ]
 [Supply: (?n1 mbi-vc:launch ?t ) (?t mbi-vc:transactsTo ?n2 )
(?n1 rdf:type mbi-vc:Supplier) (?n2 rdf:type mbi-vc:Architect) ->
( ?n1 mbi-vc:Channel ?n2)]
```

The figure 7 shows the user view model after inferencing processing.

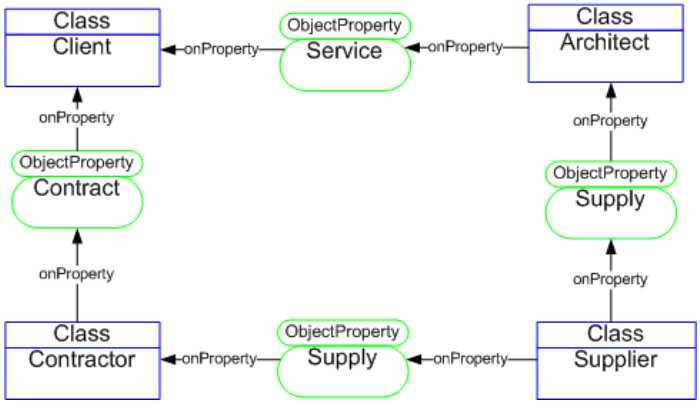

Fig. 7. Post-inference user data view

By performing the adaptation process, the user does not have to stay at the same granularity level as described in the raw data for example transactions. Instead, his attention of value-chain is more likely addressed in a customised user view over the industry domain. The outcome of view adaption is a new data model that can be well fitted into the SQIS architecture. The SPARQL [13] query is executed against use view model. The query result is visualized in a Java applet in web interface empowered by Prefuse [10], see figure 8.

Our work has shown that a general view of a network of data can be can be constrained by user preferences to show value chains within the data that are of specific interest to the user. The user preference is captured into inference rules that can process the underlying network of data to provide higher level relationships that make the users analysis and decision making much easier.

7 Conclusions

In this paper the MBI project's first prototype has been described in outline and the specific "Micro segmentation" and "Value Chains" use cases explained with real industry data. Both cases demonstrate how with Semantic Web techniques

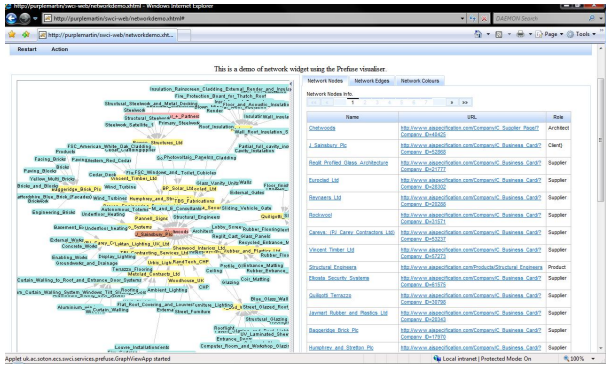

Fig. 8. Network data representation

it is possible to extract data from unstructured and semi-structured text from the Web, transform the data into RDF and integrate it with a structured data backbone. Moreover, extra value has been created by applying inference rules to the raw data pulling out new information useful in providing structured views to the end-users. The prototype has demonstrated to the consortium members the facilities Semantic Web technologies can offer when there is a need for data integration. In further work the uses cases presented will be moved to a large scale data backbone encompassing the entire UK data space of businesses and Semantic Web techniques will be included for improved modeling of "propensity to buy".

References

1. AKT research programme http://www.aktors.org (accessed on 04/2008).
2. AktivePSI http://www.aktors.org/interns/2006/aktivepsi/index.php (accessed on 03/2008).
3. The architects journal www.ajspecification.com (accessed on 04/2008).
4. http://incubator.apache.org/uima/ (accessed on 02/2008).
5. http://www.mycamden.co.uk/ (accessed on 05/2008).
6. Java server faces http://java.sun.com/javaee/javaserverfaces/ (accessed on 05/2008).
7. Ordenance survey point of interest database. http://www.ordnancesurvey.co.uk/oswebsite/products/pointsofinterest/ (accessed on 01/2008).
8. Pdfbox - java pdf library www.pdfbox.org (accessed on 04/2008).
9. PH megafile http://www.phgroup.com/mfmap.html (accessed on 04/2008).
10. The prefuse visualization toolkit http://prefuse.org/ (accessed on 05/2008).
11. Resource description framework (RDF) http://www.w3.org/rdf/ (accessed on 04/2008).
12. Simple knowledge organization system (SKOS) http://www.w3.org/2004/02/skos/ (accessed on 04/2008).

13. SPARQL query language for rdf http://www.w3.org/tr/rdf-sparql-query/ (accessed on 04/2008).
14. SWAD europe deliverable 12.1.1: Semantic web applications - analysis and selection, http://www.w3.org/2001/sw/europe/reports/ (accessed on 05/2008).
15. The UK standard industrial classification of economic activities uk sic(92) http://www.statistics.gov.uk/ (accessed on 04/2008).
16. Unstructured data: Attacking a myth, 9/5/2007, by stephen swoyer http://www.tdwi.org/news/display.aspx?id=8577 (accessed on 04/2008).
17. Web ontology language (OWL) http://www.w3.org/2004/owl/ (accessed on 04/2008).
18. WEBSPHINX: A personal, customizable web crawler http://www.cs.cmu.edu/ rcm/websphinx/ (accessed on 04/2008).
19. XHTML www.w3.org/tr/xhtml1/.
20. Harith Alani, David Dupplaw, John Sheridan, Kieron O'Hara, John Darlington, Nigel Shadbolt, and Carol Tullo. Unlocking the potential of public sector information with semantic web technology. In *ISWC/ASWC*, pages 708–721, 2007.
21. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: Implementing the semantic web recommendations. Technical Report HPL-2003., 2003.
22. F. Ciravegna, S. Chapman, A. Dingli, and Y. Wilks. Learning to harvest information for the semantic web, 2004.
23. H. Cunningham. Jape – a java annotation patterns engine. Technical Report, Department of Computer Science, University of Sheffield., 1999.
24. H. Cunningham, D. Maynard, V. Tablan, C. Ursu, and K. Bontcheva. Developing language processing components with gate, 2001.
25. Marc Ehrig and Alexander Maedche. Ontology-focused crawling of web documents. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, pages 1174–1178, New York, NY, USA, 2003. ACM.
26. Amit P. Sheth Kunal Verma Kaarthik Sivashanmugam, John A. Miller. Framework for semantic web process composition. *International Journal of Electronic Commerce*, 9, 2005.
27. Alfons Kemper and Christian Wiesner. Building scalable electronic market places using hyperquery-based distributed query processing. *World Wide Web*, 8(1):27–60, 2005.
28. Thomas Leonard and Hugh Glaser. Large scale acquisition and maintenance from the web without source access. In Siegfried Handschuh, Rose Dieng-Kuntz, and Steffan Staab, editors, *Workshop 4, Knowledge Markup and Semantic Annotation, K-CAP 2001*, pages 97–101, October 2001.
29. Eduardo Mena, Vipul Kashyap, Amit P. Sheth, and Arantza Illarramendi. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. In *Conference on Cooperative Information Systems*, pages 14–25, 1996.
30. Robert C. Miller and Krishna Bharat. SPHINX: A framework for creating personal, site-specific web crawlers. *Computer Networks*, 30(1-7):119–130, 1998.
31. Natalya Fridman Noy and Mark A. Musen. PROMPT: Algorithm and tool for automated ontology merging and alignment. In *AAAI/IAAI*, pages 450–455, 2000.
32. Bill O'Connell. Building an information on demand enterprise that integrates both operational and strategic business intelligence. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*, pages 85–86, New York, NY, USA, 2007. ACM.

33. Michael E. Porter. *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Simon and Schuster, 1980.
34. Lucia Specia and Enrico Motta. A hybrid approach for relation extraction aimed at the semantic web. In *FQAS*, pages 564–576, 2006.
35. Paul Thompson. Dynamic integration of distributed semantic services: Infrastructure for process queries and question answering. In *HLT-NAACL*, 2003.