

# SEMANTICALLY ENHANCED PASSAGE RETRIEVAL FOR BUSINESS ANALYSIS ACTIVITY

Thai, VinhTuan, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Newcastle Rd, Galway, Ireland, vinhtuan.thai@deri.org

Davis, Brian, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Newcastle Rd, Galway, Ireland, brian.davis@deri.org

O'Riain, Sean, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Newcastle Rd, Galway, Ireland, sean.oriain@deri.org

O'Sullivan, David, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Newcastle Rd, Galway, Ireland, david.osullivan@deri.org

Handschuh, Siegfried, Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, Newcastle Rd, Galway, Ireland, siegfried.handschuh@deri.org

## Abstract

*Business analysis activity often involves the identification and extraction of information from business reports. The unstructured content of business reports poses a challenge to current Business Intelligence technologies. In this paper, we present an application that provides passage retrieval capability, which has been developed for use in business analysis activity at Hewlett-Packard™ (HP). Preliminary evaluation of the application indicates that the inclusion of domain knowledge such as domain semantics and business intelligence rules plays an important part in system performance. It dramatically reduces the time that analysts spend extracting relevant informative passages within a report or business intelligence. In addition, it can find further relevant information that would have been overlooked by the analyst.*

*Keywords: Business Analysis, Passage Retrieval, Semantic Technology.*

## 1 INTRODUCTION

Global competition is forcing businesses to compete with an ever increasing number of competitors from around the world. In addition, there is an exponential increase in the amount of information generated around individual businesses through various reports. To compete, many leading international businesses deploy considerable resources in the form of business analysts for finding accurate information and deriving competitive knowledge from it. Competitive knowledge often includes: customer trends, competitor behaviour, industrial conditions and general economic trends (Golfarelli et al., 2004). To assist analysts in extracting knowledge, a number of analytical tools are available. Current Business Intelligence (BI) applications remain fundamentally driven by data mining and data warehousing techniques, which serve to monitor, identify and gather information on topics of interest. On-line analytical processing, performed on historical data, allows for report generation from which further analysis may be performed. The exploratory functionality of BI applications is limited to the structural level i.e. extracting information that has already been rigidly structured. Therefore, unstructured data, i.e. the vast majority of information contained in reports, cannot be captured or manipulated as part of the Extract, Transform and Load (ETL) process (Gartner 2003). Unstructured data, however, represents a valuable exploitable source of company financial, business and competitive information. Thus, its quick and accurate extraction offers a real business opportunity for more enhanced decision making capabilities to the business and financial analysts alike (IBM 2005).

There is a clear requirement to support business analysts with information analysis applications that can extract important information from unstructured reports. The ability to identify meaningful information in a precise and timely manner can offer a major competitive advantage. This paper presents a passage retrieval application which caters for specific information needs of analysts. The remainder of this paper is organized as follows: Section 2 introduces the business analysis process and the requirements for an application capable of supporting this activity. Related work is highlighted in Section 3. The proposed solution is then described in Section 4. The prototype implementation and evaluation are presented in Section 5. Section 6 concludes the paper and outlines future work.

## 2 BUSINESS ANALYSIS DOMAIN

In the United States, the Securities and Exchange Commission (SEC), by law, requires all businesses to submit un-audited financial statements and reports that serve to provide a continuous view of their financial position (IBM 2005). Investors and financial professionals rely heavily upon these reports to evaluate company's performance, competitive position and any investment opportunities. Hence, these reports are often viewed by the businesses themselves, as a promotional medium that presents a corporate image and are highly important in building credibility and investor confidence. Official reports are freely available for download from the SEC's EDGAR<sup>1</sup> (Electronic Data Gathering, Analysis, and Retrieval) database in HTML format. As these reports have to comply with a set of valid disclosure requirements, they represent the single most important source of information for investment decision making (Korman 1998).

The focus of this work is on the 'Form 10-Q', which represents a valuable source of both financial accounts and business statements. While business statements present the corporate viewpoint in relation to company performance, they also discuss the more qualitative aspects of the financial accounts. They are often written in a rhetorical manner with the intended goal of restricting the reader to a more positive interpretation of the information being presented. For business analysts, despite having a clear vision of the type of information they seek, traversing this semantic camouflage to the

---

<sup>1</sup> <http://www.sec.gov/edgar.shtml>

sections that contain the more interesting information remains a time consuming and error prone task (Korman 1998). Form 10-Qs are typically 70-100 pages in size and the discourse occurs as unstructured text within the report. The analyst must identify, and filter vast amounts of information in each report and only then begin the process of analysis. The process of extracting information often begins by constructing a model of the company's financial position and then searching for information of a particular type. The ability to query a report and retrieve relevant information offers the possibility to substantially remove the manual effort required. The resource saving in terms of time reduced can then be carried forward into the actual analysis phase, thus contributing to more timely and accurate analytical activities.

In this context, our goal is to design and develop an application which could automate the manual information gathering portion of the analysis activity. To achieve this goal, it is essential that the application provides the following capabilities:

- Provide a formalism that allows for the specification and encoding of domain knowledge and business intelligence rules.
- Provide a natural language interface, allowing the user to pose (factoid) queries against a business report and to be presented with relevant passages.
- Provide a visual dimension that allows passage traversal within the context of the original report.
- Provide functionality for report generation, i.e. relevant passages returned are to be aggregated and made available as a concise report.

### **3 RELATED WORK**

There exist a number of Passage Retrieval systems that offer the capability to retrieve relevant passages from unstructured contents (Tellex et al., 2003). They, however, focus on generic domains. The use of domain specific knowledge in our work allows the application to deal with business reports from the business analyst's point of view.

To enable the inclusion of specific domain knowledge into the information extraction process, Ontology-based Information Extraction (OBIE) and Semantic annotation techniques have also recently come into use in the area of Information Retrieval. A notable example is the Knowledge and Information Management (KIM) platform (Kiryakov et al., 2005), which enables the automatic semantic annotation, indexing and retrieval of texts. The core ontologies in KIM can be extended with domain knowledge for use within a particular domain. However, to our knowledge, it neither offers the passage retrieval capability for the information seeking process nor caters specifically for the business analysis task.

A related application within the context of the BI domain is the OBIE system developed within the MUSING project for the purpose of content mining (Maynard et al., 2007). It utilizes domain ontologies to extract relevant information from business reports which can then be used for analysis of financial and operational risk and in theory, within other BI applications such as Company Intelligence (Maynard et al., 2007). Similarly, the Lixto system (Baumgartner et al., 2007) is designed toward extracting information from competing businesses such as products, prices. However, these systems are not specifically designed to support business analyst engaging in the analysis of Form10-Q reports, but is rather focused on the extraction of specific types of information from free text in order to create a knowledge base. Our work applies OBIE techniques to the Business Intelligence (BI) domain. However, it is targeted toward retrieval from a text based knowledge source as opposed to extracting instances of concepts within free text and subsequently populating the ontology. Nor do we query or

reason over the populated knowledge base. This is due to the requirements specific to the business analysis activity.

Furthermore, our work differs from other Question Answering (QA) systems with regard to what constitutes an answer. Existing QA applications return precise words and phrases as answers to questions (TREC 2005). Consequently, contextual information is not returned which in fact may provide essential linkages to other important and relevant information. A QA system which is quite related to ours is the work of Zhang et al. (2004), which is built for the construction domain. In this work, the domain semantics are taken into account while measuring the similarity between questions and passages. The knowledge base in their work is a taxonomy of terminologies. In our work, however, the domain ontology encodes not only taxonomic relationships but also other types of relationship to model business intelligence rules, which play a key role during query expansion. Finally, the advent of the Semantic Web has motivated the growth of a branch of QA research focusing on answering natural language questions against formal domain ontologies. Attempto Controlled English (ACE) (Bernstein et al., 2004) and AquaLog (Lopez et al., 2005) are recent research additions to this area. These systems, unlike ours, do not target unstructured contents.

#### 4 PROPOSED SOLUTION

To meet the requirements outlined in Section 2, we propose an Information Retrieval (IR) application in which business intelligence rules are taken into consideration in finding relevant passages. Our approach starts with the modelling of domain knowledge, and then source reports and question inputs are analyzed accordingly to find answers. The system architecture is shown in Figure 1, and is followed by the descriptions of its components.

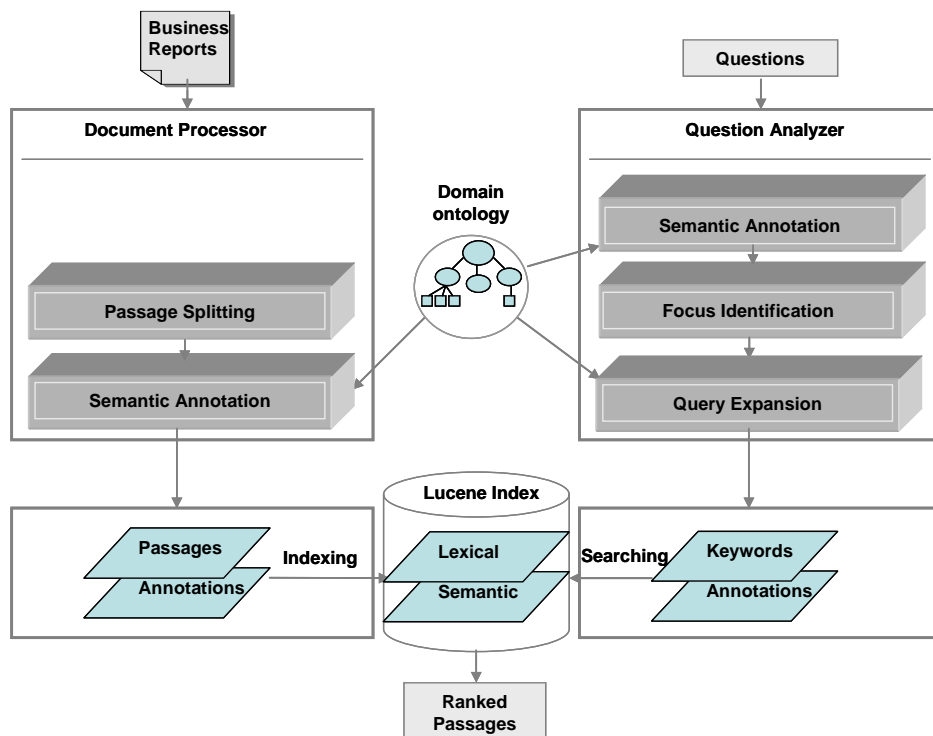


Figure 1. System Architecture

## 4.1 Domain Knowledge Modelling

Since domain knowledge is essential to the business analysis activity, various characteristics of the domain, such as its important elements and the relationships between them, need to be identified and formally represented in an ontology. Since the domain knowledge is specific to the business analysis activity, no existing ontologies in the business field are required and it is necessary to build a new ontology. We apply the ENTERPRISE methodology proposed by Uschold and King (1995) since it allows for a clear separation between building a conceptual model and its formal knowledge representation (Pinto and Martins, 2004). This approach is particularly useful as it enables the analysts, who are experts in the business field but not specialized in the knowledge representation and engineering field, to first define the conceptual model and subsequently leave the task of ontology construction to the knowledge engineers. As such, we initiated the ontology engineering process by engaging two analysts in the following activities:

- Identifying key concepts and relationships between them in the business analysis domain.
- Producing the precise, unambiguous natural language definitions for such concepts and relationships
- Identifying unambiguous terms to refer to such concepts and relationships.

Once the conceptual model was built, it was encoded into an OWL ontology using SWOOP, an ontology engineering tool (Kalyanpur et al., 2005). The unambiguous terms defining ontological concepts are in fact the terms or phrases of interest to the analysts while analyzing business reports. Of importance in the ontology are two types of relationships between domain concepts: the *taxonomic* relationships and the *relatesTo* relationships. The first type refers to the hierarchical relationships between domain concepts. The second type refers to the relationships between concepts that are meaningfully related to each other and together they help the analysts to build up certain interpretations that cannot be attained based upon these concepts alone. The use of these two types of relationships is explained in detail in Section 4.5. The next section describes how business reports are analyzed based on the ontology built.

## 4.2 Document Processing

All business reports are split into passages and entities of interest to the analyst are semantically annotated by the Document Processor.

### 4.2.1 Passage Splitting

Business reports contain both numerical data and free text in HTML format. However, it is the discourse within the free text areas that is of value to the business analyst. Therefore, it is important to identify passage boundaries for indexing and searching purposes. The HTML markups of these reports are well-formed yet complex and hence hinder the possibility of extracting passages based simply on their existing markups. Therefore, we employed language engineering techniques to identify the boundaries of passages which are then semantically annotated as described next.

### 4.2.2 Semantic Annotation

Entities of interest to the analysts are semantically annotated in order to associate them with the respective ontological concepts. This, in effect, anchors those entities to their corresponding semantic information. For instance, the term “product”, together with its synonyms such as “item” and “goods”, is associated with the concept “Product” in the domain ontology. It is also important that morphological variations of terms are accounted for in the semantic annotation process. The process is carried out via various natural language processing resources provided by the GATE library (Cunningham et al., 2002). Of importance among them are the Gazetteers which perform gazetteer list lookup to link the recognized entities to concepts in the ontology. The mapping list is based upon the

conceptual model defined by the domain experts. However, there are also cases whereby concepts are represented by complex phrases following different patterns. The Gazetteers cannot carry out the annotation process in these cases and therefore JAPE (Java Annotation Patterns Engine) rules are used instead. The identified passages and their associated annotations are then stored into an index.

### 4.3 Indexing

Apache Lucene<sup>3</sup> is used as the index and search engine. The index structure consists of three fields: “Id”, “Contents”, and “Annotations”. The “Contents” field represents the *lexical* layer and the “Annotations” represents the *semantic* layer. These two layers enable the matching of questions and passages not only by exact words, but also by the domain semantics.

### 4.4 Question Analysis

Natural language questions input by the users are semantically annotated in the same manner as the passages extracted from business reports. In addition, question focus identification is performed to find out the most important element in the question, which must be present in any returned passages. For example, questions starting with “*What are the strategies .. ?*” ask about strategies so their focus is the concept “Strategy”. Question focuses, however, do not always correspond to domain concepts. Therefore, it is important that only passages containing a question’s focus or its annotations should be considered as potentially relevant passages.

The identification of question focus is based upon the syntactic structure of questions. Analysis of the set of questions posed by the analysts indicates that there are three patterns of questions:

- “*What is/are the <noun> ..?* ” or “*What <noun> is/are...?* ”
- “*Is/Are there any <noun> ...?* ”
- “*Have/Has the ...?* ”

Among the above three patterns, question focus can only be identified in questions belonging to the first two, not in the third one as its sentence structure is complex. Furthermore, since semantic annotations are used in Lucene queries, the words or phrases annotated by them are removed from the question itself to avoid duplicated queries.

Each concept annotated in a question is expanded with concepts that have either of the following two relationships:

- **Taxonomic** relationship: annotated concepts are expanded to include all their sub-concepts. For example, if “Legal” is one of the concepts annotated in the question, then its sub-concepts such as “Litigation” and “Regulation” are also taken into consideration.
- **relatesTo** relationship: This is an important relationship since it represents the links between one concept and those that are meaningfully related to it in the Business Analysis domain. For example, the concept “Change” can be associated with, among others, the concept “SalesOfAssets” and therefore the latter is also included in the query.

### 4.5 Searching

Each question is translated into a Boolean query based on its modified question string together with its set of expanded annotations. Keywords parsed from the modified question string are used to query on the “Contents” field and annotations are used to query on the “Annotations” field. On the one hand, since the above-mentioned keywords are not associated with any important concepts in the domain ontology, potentially relevant passages may or may not contain them. There is, however, an exception

---

<sup>3</sup> <http://lucene.apache.org/>

in the case of question focus which is not annotated. Since question focus is the most important element in a question, it needs to appear within any returned passages. On the other hand, since annotations represent important concepts in the domain, their appearances in potentially relevant passages are mandatory. Expanded annotations of the original annotations are also included but their appearances are optional as they belong to the same category as their parent concept.

To illustrate the above query formulation strategy, take for example the question “*Are any changes planned for the current business model?*”. Its question focus is the concept “Change”, its annotation set includes “Plan” and “BusinessModel”, and its modified string is “current” once all the stop-words and annotated terms are removed. Query expansion is then carried out. Although none of these concept annotations have sub-concepts in the domain ontology, the concept Change has *relatesTo* relationship with, among others, the concept “SaleOfAssets”. Hence, a Boolean query is formulated as follows:

(contents:current) +(annotations:Change annotations:SaleOfAssets)  
+(annotations:Plan) +(annotations:BusinessModel)

This query formulation, in effect, allows for querying, at the *lexical* layer, of terms in a question against the surface texts in the collection of extracted passages and also at the *semantic* layer - of the expanded annotation sets in a question against the annotations of the extracted passages. Given the rhetorical language used in reports, the use of domain knowledge to annotate business reports as well as questions and to expand query terms in the semantic layer demonstrates the beneficial use of semantic technology in that:

- The availability of semantic annotations, in effect, helps to bridge the gap between the words used in questions and those used in the texts. In many cases, the words used by analysts in the questions can be expressed differently from the ones written in business reports, even though they have the same meaning. For instance, when appearances of “item” in a certain passage are associated with the concept “Product”, they provide an abstraction that might help to answer such question as “*Which products did the competitors release?*”. In this case, the term “products” in the question is also annotated with the concept “Product”, therefore, a match can be found even though the question and the passage use different words to refer to the same concept.
- Query expansion based on the *taxonomic* relationships in the domain knowledge helps to enrich the query with terms that belong to the same category of the original query terms.
- Query expansion based on the *relatesTo* relationships models the way domain experts find answers for questions. In fact, this specific query expansion strategy is the direct application of business intelligence rules to find information of interest, which cannot be found based simply upon keywords and their respective annotations.

## 5 PROTOTYPE IMPLEMENTATION AND EVALUATION

### 5.1 Prototype Implementation

A screenshot of the prototype is shown in Figure 2. The contents of Form 10-Q are displayed in the Report Viewer panel. Once a report is processed the analysts can pose questions in natural language. The questions are analyzed and relevant passages are presented. The analysts can quickly traverse through the returned passages, and evaluate them for relevant contents. Returned passages are highlighted and presented in order of relevance to a question. Therefore, the application guides the analyst to the most relevant passage first regardless of its location in the report. In the next section, we describe the evaluation and discuss our findings as well as some initial user feedback.

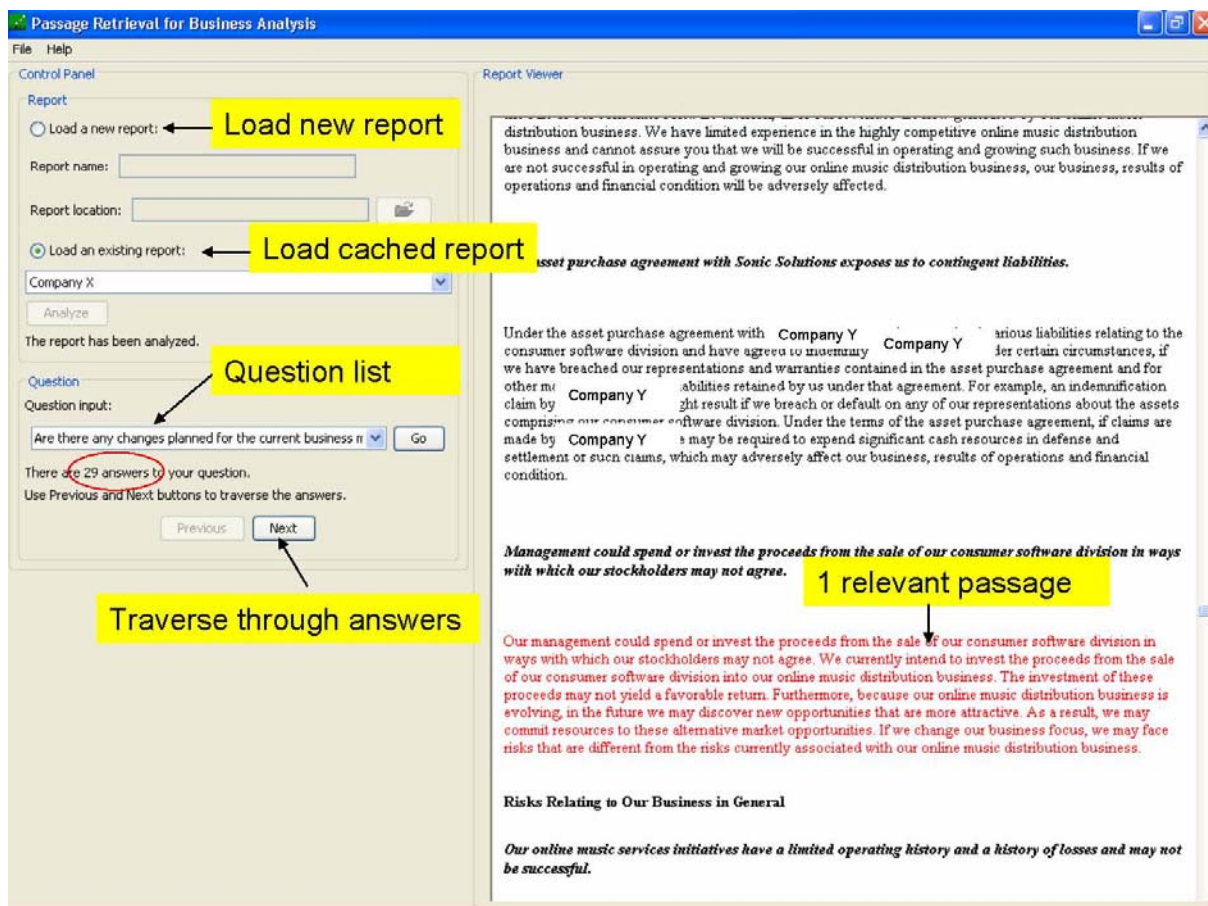


Figure 2. Passage Retrieval Prototype Screenshot

## 5.2 Evaluation

Since this system is designed and built specifically for the domain of business analysis, it is unique with regard to source documents, domain knowledge, and information needs. Consequently, no previous reference test set could be re-used for evaluation purposes. Therefore, it was necessary to build our own reference collection. The process of test set creation is, however, time-consuming and costly as analysts are scarce in number and constrained by the strict time-lines inherent to the private sector. Therefore, we choose to carry out an evaluation involving one domain expert, who could judge the relevance of returned passages as answers to a set of test questions. The evaluation setup is described next.

### 5.2.1 Evaluation Setup

Our test collection consists of the following:

- A set of 3 business reports (3 different Form 10-Q filings downloaded from EDGAR)
- A set of 10 questions<sup>4</sup> posed by the analyst.
- A set of relevant judgments.

In IR evaluation, a relevance judgment normally is a binary classification indicating relevant or not relevant. However, in this domain, a returned passage may be judged by the analyst as (1) relevant or

<sup>4</sup> Accessible at: <http://sw.derl.org/svn/sw/2006/10/QA/TestQuestions/questions.txt>



(2) irrelevant or (3) somewhat relevant (informative) with respect to the given question. Allowing for the existence of this category caters to useful background or context information (i.e. it is partially relevant or may prove of interest to the analyst). Thus, we opted for a lenient evaluation scheme whereby somewhat relevant (informative) answers were treated as relevant. Furthermore, due to the information overload issue, the business analyst may often overlook passages that qualify as answers while performing the task manually. Therefore, the gold standard set of relevance judgments was formed by the analyst’s judgments on the combination of answers manually marked up by himself and those automatically suggested by the application for each question.

In addition, we are also interested in the impact of domain knowledge on the application’s performance. Therefore, we also built a baseline Passage Retrieval system that does not make use of the domain ontology for the purpose of semantic annotation and query expansion. In other words, the baseline system only uses exact word matching to retrieve relevant passages for each question. This baseline system was evaluated using the same procedure as with the full-fledged system.

Finally, we measured the percentage time saved as a consequence of using the application by asking the analyst to keep track of the time spent retrieving passages both manually and automatically using the tool.

Since for each report-question pair, the returned answer set may have consisted of multiple passages we were only interested in retrieval performance at rank 10, i.e. we only considered the relevance of at most the first 10 of all returned passages. Therefore, for comparison between performances of the analyst and both of the above-mentioned systems, we used in our evaluation the following metrics: Recall at rank 10, Precision at rank 10, and Percentage of time saved.

### 5.2.2 Results

The retrieval results are shown in Table 1.

	Business analyst	Full-fledged system	Baseline system
Average recall value	0.646	0.542	0.294
Average precision value	1	0.578	0.25

*Table 1. Preliminary evaluation results*

In terms of timing, on average the analyst took 180 minutes to manually go through each report. The first 120 minutes were consumed reading the report and identifying noteworthy passages. The remaining 60 minutes were taken up for inferring answers based on the manual markups. By using the QA system, it took the analyst 3 minutes to analyze a report and 60 minutes to infer judgments regarding the correctness of the given answers based on the passages retrieved.

### 5.2.3 Discussion

Owing to the large cost associated with business resources limiting our evaluator pool to one and restricting our report test corpus, we refrain from drawing conclusive remarks on the recall and precision values as they are not statistically significant, no matter how tempting. However, we believe that the retrieval results considered together with user feedback from the analyst still provide valid insight into the prototype’s performance and merit.

In relation to retrieval results, the average recall and precision values of the full-fledged system were 0.542 and 0.578 respectively. Though this indicates that its overall retrieval performance was lower than that of the domain expert, there are a number of noteworthy observations. For instance, for all reports, the analyst identified correct sets of answers for 7 out of the 30 test question/report pairs. The

recall values of less than 1 on the other 23 question-report pairs indicate that some passages were actually overlooked by the analyst. This reflects the fact that the vast amount of qualitative and quantitative information contained within business reports remains a challenge for the manual identification and extraction of information from them. It should also be noted that the prototype achieved better recall values in 11 question-report pairs, which means that the analyst missed out more answers than the prototype in 37% of the question-report pairs.

Another observation is that the domain knowledge represented by the domain ontology played an important role in finding the right answers. The baseline system, which is based solely on the similarity between terms mentions contained within the questions and those in the test document and not on domain semantics, achieved lower recall and precision values than those of the analyst and the full-fledged system. This may be attributed to the fact that the language used in business reports varies greatly. As such, if there is no mechanism to include synonyms and related concepts of queries' terms, the baseline system will be unable to meet the analysts' information needs.

Based on the business analyst's feedback, when using the application, the analyst obtained a significant resource saving on the manual analysis of Form10-Q. The time saving represents a saving of 65% on the overall resource requirement. With analysis resources representing a substantial cost, these time-saving and reduced manual effort are of particular interest to the analyst community.

In addition, the analyst also noted the advantage of using the automated tool in terms of visual navigation over the answer set returned for each question. The tool enables the traversal of returned passages in accordance with their relevance to the input question. As a result, by using the tool, the analyst is guided to the most relevant passages first regardless of their order of appearance within the report. This is of interest to the analysts since without this navigation feature, currently they would have to manually analyze reports from beginning until end in order to find relevant answers. Finally, the analyst has stated that the current prototype would be particularly helpful in providing aid for the *ab initio* or novice business analyst. While the evaluation is preliminary, the initial findings affirm the usefulness of providing automated tools to support business analysis activity.

## 6 CONCLUSIONS

In this paper, we have presented work targeting at assisting business analysts in dealing with the information overload problem while analyzing Form 10-Q reports in order to identify new business opportunities. The current process is time consuming, error-prone and requires a large number of business analysts. To support the business analysts, we proposed a Passage Retrieval application which presents them with relevant information in a timely manner. Of importance within the proposed approach is the use of, among others, semantic technology. Semantic technology enables the modelling of domain knowledge and the inclusion of business intelligence rules into the information identification and extraction process.

We conducted a preliminary evaluation of the system based on its retrieval performance and user feedback. The results are favourable and benefit substantially from the inclusion of domain semantics within the system. Preliminary feedback indicates that a fully deployed system could save significant time and resources. Based on the user feedback, the prototype presented in this paper is to be extended with more features. We are currently adding the off-line or batch mode processing feature to the tool, thus enabling the analysts to delegate the extraction process to non-experts through the provision of a selection of predefined questions for a given set of reports. Future work will also involve identifying question patterns for which the system cannot achieve good recall and precise values. This will require a more extensive evaluation involving a larger collection of business reports and many more test questions. Moreover, it is also noted by the analyst that the application prototype returns as answers some passages that contain generic statements regarding regulations rather than specific company

information. Resolving this issue will help to eliminate these false positives and hence improve the retrieval performance.

## Acknowledgments

We would like to thank Michael Turley for participating in the preliminary evaluation. This work is supported by the Science Foundation Ireland (SFI) under the DERI-Lion project (SFI/02/CE1/1131) and partially by the European Commission 6th Framework Programme in context of the EU IST NEPOMUK IP - The Social Semantic Desktop Project, FP6-027705.

## References

- Baumgartner, R., Frölich, O., & Gottlob, G. (2007). The Lixto systems applications in Business Intelligence and Semantic Web, in: ESWC'07: Proceedings of the European Semantic Web Conference, Innsbruck, Austria, pp. 16-26.
- Bernstein, A., Kaufmann, E., Fuchs, N. E., & von Bonin, J. (2004). Talking to the semantic web a controlled English query interface for ontologies, in: 14th Workshop on Information Technology and Systems, pp. 212-217.
- Cunningham, H., Maynard, D., Bontcheva, K., & Tablan., V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, in: ACL'02: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, Philadelphia.
- Gartner (2003). Business intelligence tools: Perspective, Gartner Research.
- Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond data warehousing: what's next in business intelligence?, in: DOLAP '04: Proceedings of the 7th ACM international workshop on Data warehousing and OLAP, ACM Press, New York, NY, USA, pp. 1-6.
- IBM (2005). Guide to understanding basic financials, Available online at: <http://www.ibm.com/annualreport/2005/guide.shtml> ( Last access date: 30/07/2007).
- Kalyanpur, A., Parsia, B., Sirin, E., Cuenca-Grau, B., &Hendler, J. (2005). SWOOP: A 'Web' Ontology Editing Browser, Journal of Web Semantics, Vol 4(2).
- Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2005). Semantic annotation, indexing, and retrieval, Journal of Web Semantics 2, 39.
- Korman, R. (1998). Investing it; mining for nuggets of financial data, The New York Times (June 21 1998).
- Lopez, V., Pasin, M., & Motta, E. (2005). AquaLog: An ontology-portable question answering system for the semantic web, in: ESWC, pp. 546-562.
- Maynard, D., Saggion, H., Yankova, M., Bontcheva, K., & Peters, W. (2007). Natural language technology for information integration in business intelligence, in: W. Abramowicz (Ed.), 10th International Conference on Business Information Systems, Poland.
- Pinto, H. S., & Martins, J. P. (2004). Ontologies: How can they be built?, Knowledge and Information Systems 6 (4), 441-464.
- Tellex, S., Katz, B., Lin, J., Fernandes, A., & Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering, in: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM Press, New York, NY,USA, pp. 41-47.
- TREC (2005). Text retrieval conference TREC. <http://trec.nist.gov>
- Uschold, M. & King, M. (1995). Towards a methodology for building ontologies, in: IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canada.
- Zhang, Z., Sylva, L. D., Davidson, C., Lizarralde, G., & Nie, J.-Y. (2004). Domain-specific QA for construction sector, in: Proceedings of SIGIR 04 Workshop: Information Retrieval for Question Answering.