A SEAMLESS QUERY APPROACH TO MULTIPLE DATABASES

ZUBAIRU, H. A[1]; OYEFOLAHAN, I. O[2]; BABAKANO, F. J[3]; & ETUK, S. O[4].
Department of Information and Media Technology,
Federal University of Technology, Minna, Nigeria
E-mail: abu.zubairu@futminna.edu.ng; o.ishaq@futminna.edu.ng;
faiza.bkano@futminna.edu.ng; abiolastella@futminna.edu.ng
Phone No: +234-803-658-4777

Abstract
The maturity of Internet has given rises to many e-commerce sites that users may need to access. Therefore, there is need to provide users with easy and flexible access to information from multiple, dispersed and heterogeneous e-commerce data sources via single point of access. E-commerce sites are continuously emerging, maintained by different organizations and managed independently. The combination of data from different e-commerce sites on the internet usually fails due to syntactic and semantic differences. The access, retrieval and utilization of information from the different e-commerce web portal necessitate the need for easy and seamless access to several databases. Semantic difference technically refers to as semantic heterogeneity is a major issue that needs to be resolve in order to create interoperability among several databases on the internet. Though, there are existing approaches, but they are no longer efficient in the current reality of ubiquitous data on the internet. In this paper, an architecture for unify interface to multiple data sources on the internet is proposed. The paper adopts ontology-based approach using WordNet as a semantic dictionary for the reconciliation of semantic conflicts between the concepts or terms. Experimental evaluation of the proposed approach shows that the synonyms of a queried term are generated as hit from the available databases, which implies the feasibility and effectiveness of the approach.

Keywords: Ontology, WordNet, Query Reformulation, Multiple Databases

Introduction

The revolution in the 21st century has made Internet an alternative medium for the exchange of goods and services. One aspect of human survival that is becoming dependent on the Internet is commerce. Internet enabled commerce popularly known as e-commerce is the act of selling goods online using computers or devices that allow online transactions.

As e-commerce is widely embraced and computer is becoming more affordable, databases become more widely used. Exchange of data between multiple and heterogeneous distributed databases become necessary, and consequently database interoperability and integration have been an open research area. The e-commerce domain has necessitated a new requirement for information services that are homogenous in their presentation, open in terms of software architecture and a global scope (Antoniou & Harmelen, 2008). There is need to support smooth interaction within the different, independent, distributed data sources of same domain running on heterogeneous platforms.

To address this issue, several schemes were developed; global schema approach (Asfari, Doan, & Sansonnet, 2009), which defines a global schema over the component database systems that capture the union of the information content of the component schemas, Federated database approach (Ghawi & Cullot, 2009), which exports schemas of distributed database and integrates with the local schema to provide the necessary views for the local users and Ontology-based approach (Sharma & Gulati, 2010), which provides explicit specification for entities and ontology languages for querying the different data sources.

Each of the schemes is tailored toward addressing a specific challenge in multi-database system. While global schema and federated database approaches focuses on syntactic issues, ontology based focus on semantics. Ontology which is a formal and explicit specification of a shared conceptualization, Gruber (2009), will play a crucial role in describing the contents of different data sources and facilitating their interoperability. However, technology evolution is rapidly growing, the use of global ontology and the require conformity from different data sources to that ontology as currently the case, is no longer efficient in a real multi-database environment.

This research presents a different approach to query multiple databases using ontology. The paper assumed different relational databases of same domain; storing semantically related information in syntactically different terms, unlike the existing ontology approach, there is no need for conformity to a particular ontology, but only a sharing of terms for common elements using WordNet.

This paper presented an ontology-based approach for creating a unify access to multiple and disperse database in the domain of e-commerce. The developed framework provides seamless access to multiple databases without any transformation to the existing system.
A query reformulation algorithm based on WordNet synonyms is proposed for inclusion of semantically related concepts. The rest of the paper is organized as follows; Section 2 review related work and WordNet ontology. Section 3 presents the problem definition. Section 4 is the methodology and query reformulation algorithm. Section 5 presents the experimental evaluation, results and discussion, findings of the research. Lastly, conclusion and future work is in section 6.

Related Work

An emerging need to access information from different databases has resulted development of some prototypes (Ghawi & Cullot, 2009; Guarino & Giaretta, 1995; Raji Ghawi. 2010; Suwanmanee, Benslimane, Champin & Thiran, 2005). The term data integration used in association with accessing several databases in literatures is somewhat misleading; querying multiple databases might be a more appropriate term, since practically there is no integration. In literature, there are two basic types of data integration approaches (Halevy, 2001); global as view (GAV) systems and local as view (LAV) systems. Their difference lies in how the mappings are defined between the global view and the source. Though schema matching systems (Hanzala, Abid Ali & Fareed, 2008) uses semi-automatic mapping discovery between the global view and the sources, they have little or no success in constructing the global view itself. Effort is being made towards using reference ontology (Ghawi and Cullot, 2009), global ontology construction and its evolution remains an open challenge.

Data integration systems (Lenzerini, 2002) require a global view to be constructed before local sources are integrated into the system. The bottleneck in the integration process is the construction of the local ontologies for each data sources, inter-ontology mapping and mapping to global ontology. Current ontology approaches to database integration assume a well perfect ontology for each data sources and global level knowledge domain to build the integrated system. In reality, defining a perfect ontology for each database entity is not feasible, automatic ontology construction is still at infancy and integrators have limited knowledge of the data sources. Data sources must be integrated quickly and the scale of integration in reality makes defining a perfect ontology a challenge.

Using ontology for integration has been used previously in (Ghawi & Cullot, 2009; Guarino & Giaretta, 1995; Suwanmanee, Benslimane, Champin & Thiran, 2005). These systems require sources to completely commit to a particular ontology language and manually map all of their data to the ontology. Since ontologies have more powerful modeling constructs, their construction is more challenging. Inter-ontology mapping and mapping to the global ontology is a daunting task. It is still a challenge to construct a well-structured and perfect ontology for all entities in the databases and refine a global ontology.

Many researches have been done in database integration using ontology approach (Ghawi & Cullot, 2009; Guarino & Giaretta, 1995; Suwanmanee, Benslimane, Champin & Thiran, 2005). These researches assumed the databases involve to have a well-structured ontology. However, the world is dynamic hence; the feasibility of structuring a perfect ontology including the relations of all the resources in the various databases is highly impracticable.

Structuring a perfect ontology for all the databases that may be involve in data integration system is not feasible. In order to overcome the above difficulty, this paper proposed an approach of database integration that considers the databases in their original form without transforming them to conform to a particular ontology. To the best of our knowledge there is no search engine yet that has the capability to returned results according to their semantics or user intention. An interesting attempt can be trace in (Vivisimo, 2002), however; the detail of the system is not exposed.

This paper present an approach that uses WordNet as a controlled vocabulary, to discover possible semantics that the query inserted could be possibly implies.

WordNet

WordNet (Princeton, 2002) is an online lexical resource for English language that classify synonym terms into synsets, each expressing a distinct concept. The WordNet database contains information on semantic relatedness for words, including relations such as synonymy (same meaning), antonymy (opposite in meanings), hyponymy (hierarchical relationship), and meronymy (part-of relationship). WordNet labels the semantic relations among the words and provides explicit pattern other than meaning similarity, Patwardhan, and Pedersen (2006).

WordNet ontology is a good resource for mapping concepts. We extend the user query with the synonyms in WordNet belonging to the synset of each term contained in the user query. Though, WordNet contains different relations, however, we only exploit the synonyms part of the WordNet. WordNet ontology is used to generate semantically meaningful queries by deriving the synonyms related to the query terms.

Problem Definition

In a multi-database (integrated) system of a particular domain, the same entity may be stored with different names in different databases. Retrieving information from an integrated system, matches the user's query with available databases, through syntax; this is due to the fact that the traditional method of processing query depends on syntax. Since, the conventional query processing is sensitive to vocabulary, it is possible that the user's initial keywords do not get results of interest, but the relevant documents uses different term from the original query that are semantically equivalent to the user's term that capture user's intention.

For example, a synonym of a term found in the system, may be used in the query (TV and television, price and rate). Retrieving record from an integrated system often fail, because of syntactic and semantic differences. Consider Table 1:

Table 1: A sample relational database table

Id	Product name	price
1	Cellular phone	5000
2	Television	20000
3	Sound system	5000
4	Personal computer	3000

A user query for a "TV", from Table 1 return null, because, the tradition method of query processing is vocabulary sensitive, though, "TV" is implied by other term ("Television") which should be considered in processing the query.

Retrieving records that are semantically related to a target request will yield better results in multi-database system that may have different records describing the same entity (Amshakala & Nedunchezhian, 2011). This requirement is far beyond the capabilities of the conventional syntax based query processing. If semantic relation like synonyms could be included in the search, then semantically related terms may be retrieved. By replacing the query terms with a mechanism which can identify semantically similar terms, challenge posed by the use of different terms to define the same concept may be handled. These form the basis of the paper methodology.

Methodology

The paper adopts an approach that combines WordNet ontology as a semantic dictionary and query reformulation algorithm. The major idea is that data integration can be achieved if naming conflicts can be eliminate or resolved. Adopt semantic lexicon (WordNet) capable of determining synonyms and performing query reformulation help to resolve synonyms problems. A classical substitution principle is the underline approach. Given a search term, the paper developed a mechanism to find related terms and search for those related terms from WordNet. This is a complex task; since the system is syntax based. Therefore, the paper define some concepts that will be useful for in this approach.

Define $S = \{s_1, s_2, s_3, \text{----------} s_n\}$, to be the set of synonyms retrieved from the WordNet, where $s_n \varepsilon S$. It is a mapping from word phrases (S) to synset (s_n), that is $S \rightarrow s_n$. In particular, the mapping $S \rightarrow s_n$, connotes the following: S denotes a meaning of the (word) phrase s_k. For example: Television → {"TV", "Telly", "television", "telecast"}. This denotes that the word "television" and the synset "TV", "Telly", "television", "telecast", are semantically equivalent.

From the problem definition, it was asserts that, user may likely not enter exact term that matches the answer syntactically, however, the user(s) should be able to use or enter term or phrase that are semantically similar to the right answer. Thus, there is conceptual definition of how terms can be semantically related to one another. The set of synonyms are retrieved from WordNet (Jayaprabha & Somasundaram, 2011) database. The retrieved synonyms are ranked base on semantic similarity measurement. This was implemented using WordNet::Similarity (Patwardhan & Pedersen, 2006) that contains an algorithm for measuring semantic similarity.

Internally, WordNet uses Java WordNet search application interface (JAWS), a Java API to WordNet, to access the database. The system utilizes JAWS to retrieve and cluster the

synonyms. A query reformulation algorithm was developed that utilizes the retrieved synonyms to reformulate the user query based on the available synonyms and then generates hits from the database(s).

The architecture of the proposed system is shown in Figure 1. The user submit query and receive the output via the user interface and output panel respectively. The WordNet is far from being complete, therefore, the paper makes effort to extend the database to accommodate new terms; user can insert new terms vie user interface for inserting new term and the update WordNet module is responsible to update the WordNet database. The data source houses all the different databases in the data source. The databases in this approach are not physically present, but the system uses their addresses registered with the application to access and querying them remotely.

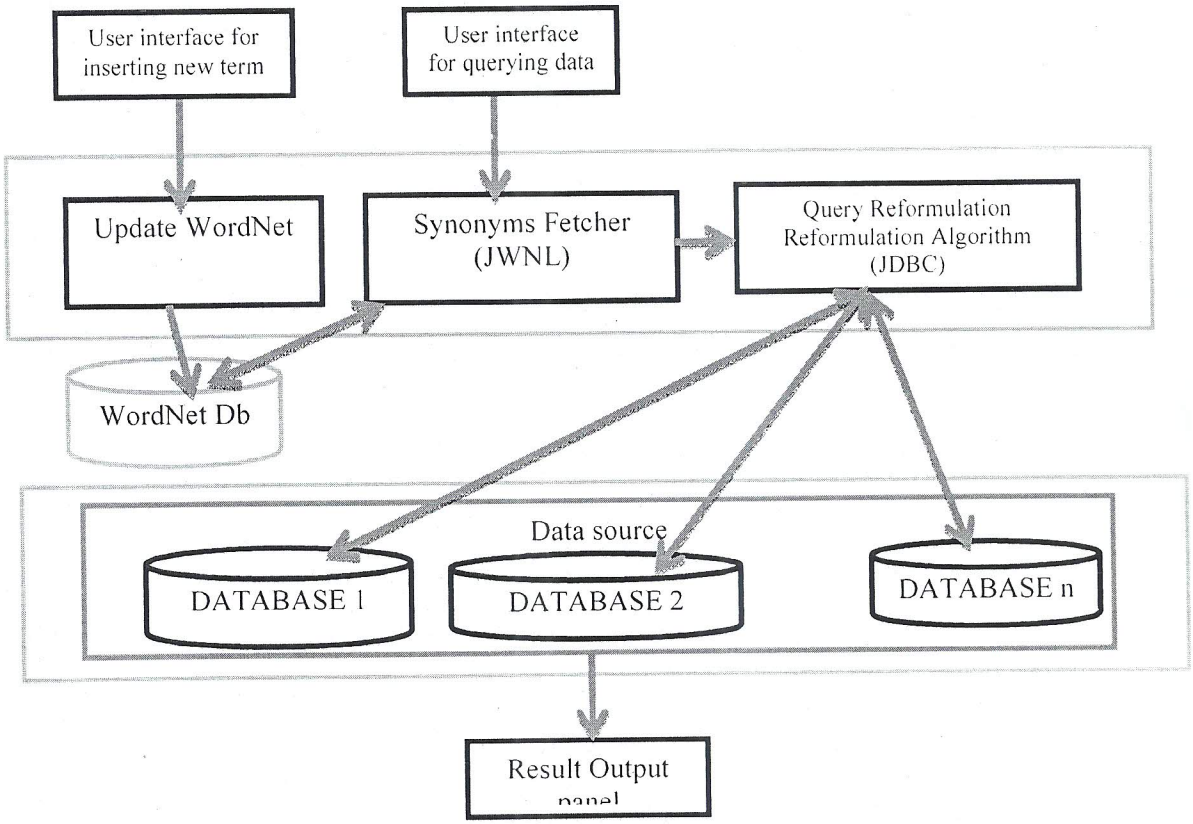

Figure 1: The System Architecture

Query reformulation

Query reformulation is also known as query expansion. Query expansion is the process of adding relevant terms to the original query (Asfari & Sansonnet, 2009). Query reformulation is a way of reformulating the user submits query based on the available synonyms from the WordNet. It is used to answer user's queries in more efficient manner by reformulating user submitted queries into semantically related concepts or synonyms. The reformulated query is called the semantic query (Ali & Khan, 2005). These semantic query is formed by using WordNet ontology support, which provide synonymous terms equivalent to keyword query term.

We propose a query reformulation algorithm, which our system will use in reformulating the input query based on the available retrieved synonyms. The algorithm is embedded in the

query reformulation module of our system. This is done in accordance to the following algorithm:

Define S = {s_1, s_2, s_3,----------s_n}, to be the set of synonyms retrieved from WordNet.
Define T = {t_1, t_2, t_3,----------t_n} to be the set of tables available in the system.
Define F = {f_1,f_2, f_3,----------f_n} to be set of field in table t_i
Result = dataset of possible output.
For each $t_i \varepsilon$ T // table t_i in T, i = 1 to n.
For each $f_j \varepsilon$ F // field f_j in T, j = 1 to n.
For each $s_k \varepsilon$ S // s_k is a subset of S, where S is a set of synonyms, and k= 1 to n
Result = Result u SELECT * from T, where f_j contains s_k
End for
End for
End for
Return result.

Experimental Evaluation

The evaluation of the concept was carried out by implementing different databases using tables presented in Amshakala and Nedunchezhian (2011). The table is a categorization of terms and their synonyms, its domain is broad enough to be realistic, and the content of the table is applicable in the domain of e-commerce and is understandable by non-experts. Table 2, table 3 and table 4, are different tables in the application data sources.

Table 2: Item table from mysql database

```
mysql> use mysqldb;
Database changed
mysql> select * from item;
+------+------------------+-------+
| id   | product_name     | price |
+------+------------------+-------+
| 1    | cellular phone   | 5000  |
| 2    | television       | 20000 |
| 3    | sound system     | 5000  |
| 4    | personal computer | 30000 |
+------+------------------+-------+
4 rows in set (0.06 sec)
```

Table 3: Commodity table from derby database

ID	PRODUCT_NAME	PRICE
1	mobile	3000
2	telecast	25000
3	music	4000
4	computer	25000

Table 4: Product table from sqlite database

id	product_name	price
1	mobile phone	4000
2	tv	25000
3	stereo	5000
4	pc	20000
5	washer	10000

Results and Discussion

The framework is evaluated by implementing a database where terms and its possible synonyms are stored in different tables. The main interface to the system is shown in figure 7.

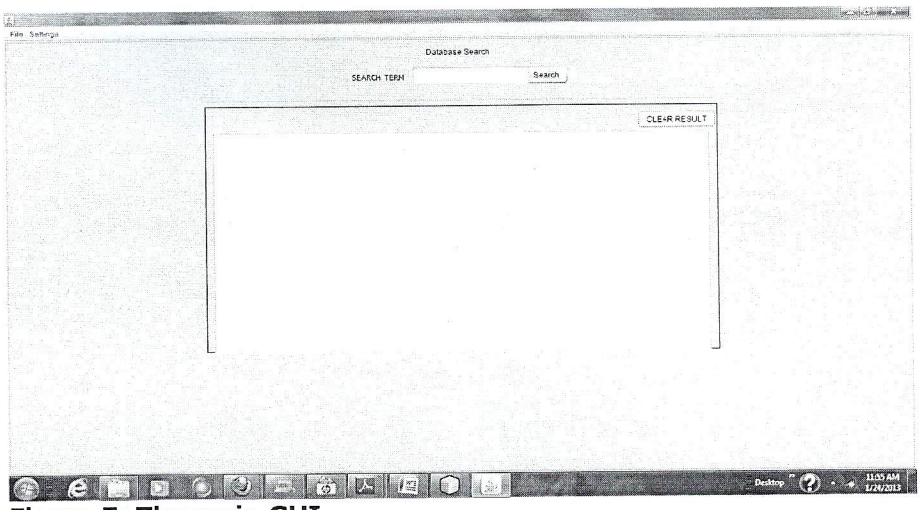

Figure 7: The main GUI

Figure 7, shows the main GUI developed using Java swing. It has two menu; file menu and settings menu. The file menu has one submenu-the exit menu. In addition, the GUI has a textbox that allows the user to search for term and a result panel to view the result.

The first evaluation was done without the semantic dictionary (WordNet), and it was observed that if user specifies a term which is not present in the database, it gives no hit. Although, the intended concepts or entity may be available. Figure 8 is the screen shot of no hit.

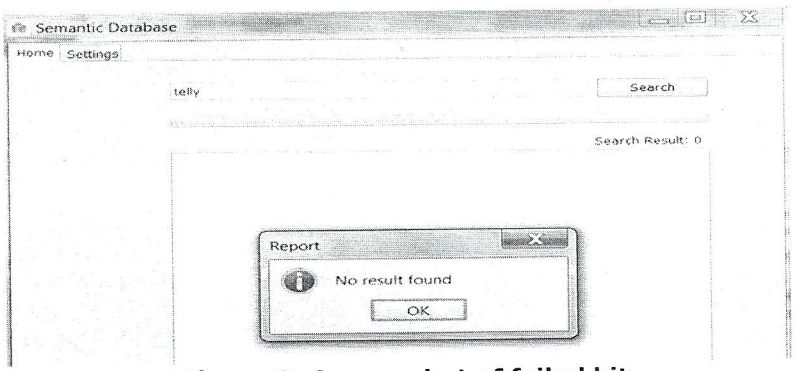

Figure 8: Screen shot of failed hit

The subsequent evaluation was done with the semantic dictionary fully incorporated and integrated into the database. A query was executed on the database and obtains the concepts or entities that match with the user's request. The system not only discovers the concept or entities that syntactically match with that of the request, it also retrieves concepts or entities that are synonymous when the exact match is not found.

For instance, if a user requests a music system, washing machine and television, the system retrieves stereo, washer and telly respectively, as shown in Figure 9, 10 and 11, respectively.

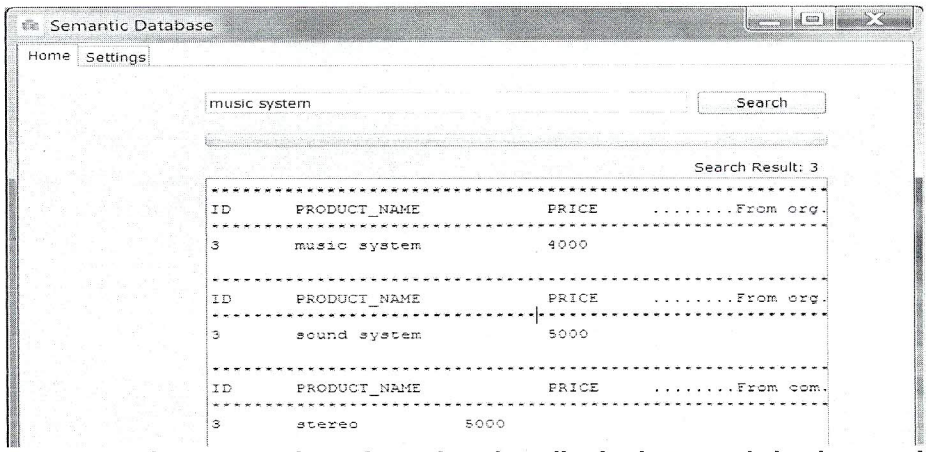

Figure 9: The system interface showing displaying result in the result panel

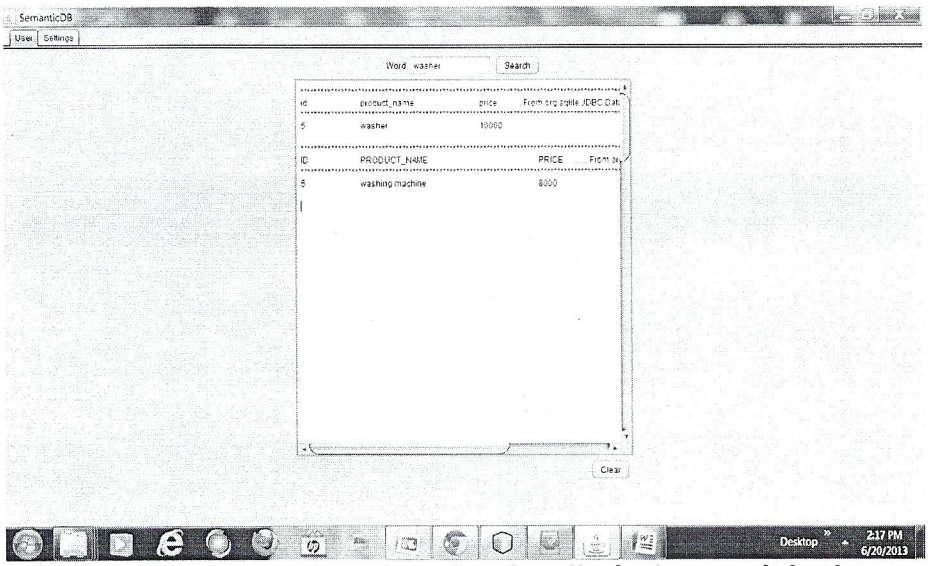

Figure 10: The system interface showing displaying result in the result panel

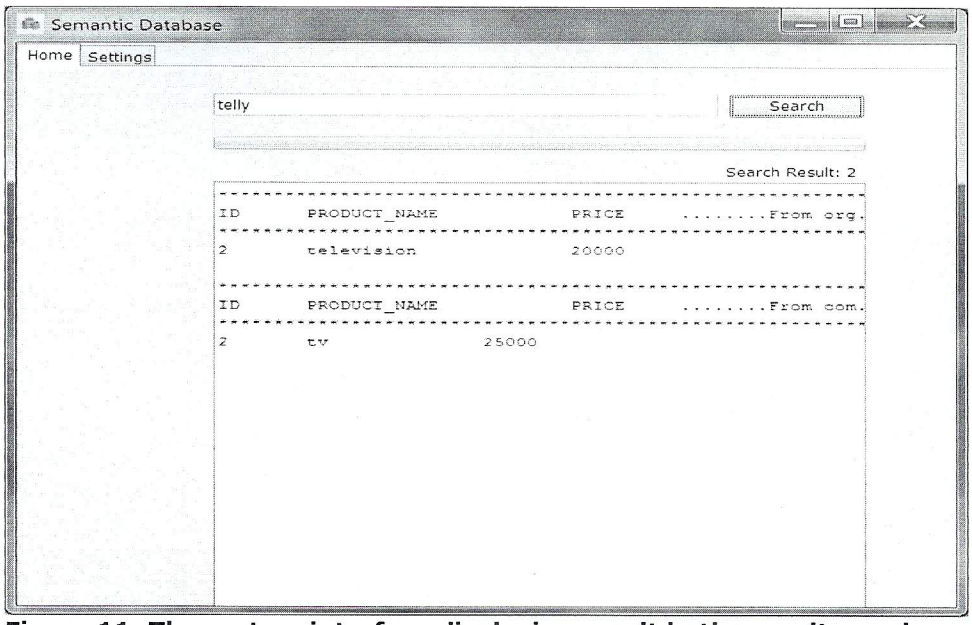

Figure 11: The system interface displaying result in the result panel

As observed from the figures above, though some instance where the term requested by the user is not available in the database, the system retrieves more semantically matching results. The system use the semantic dictionary (WordNet) to infers that music system, washer and television are all synonymous to stereo, washing machine and television respectively. The above results show an enhancement in the ability of retrieving concept based not only on exact meaning, but also on the existence of semantic relationships between the label terms.

The evaluation of the proposed approach with respect to syntax based search was performed in term of recall. For the comparison of the results, the evaluation was carried out on a system with 2.40GHz processor, 2GB RAM and on windows 7 operating system. Queries are executed on the system with and without the semantic dictionary using tables (1 - 4).

The use of semantic dictionary as proposed in this paper increase the recall of the user submitted queries compared to the syntax based query without semantic knowledge. The inclusion of semantically related terms has led to increase in the results and therefore more results are presented to the users. As the graph shown in Figure 1, the proposed approach considerably increase the user's recall.

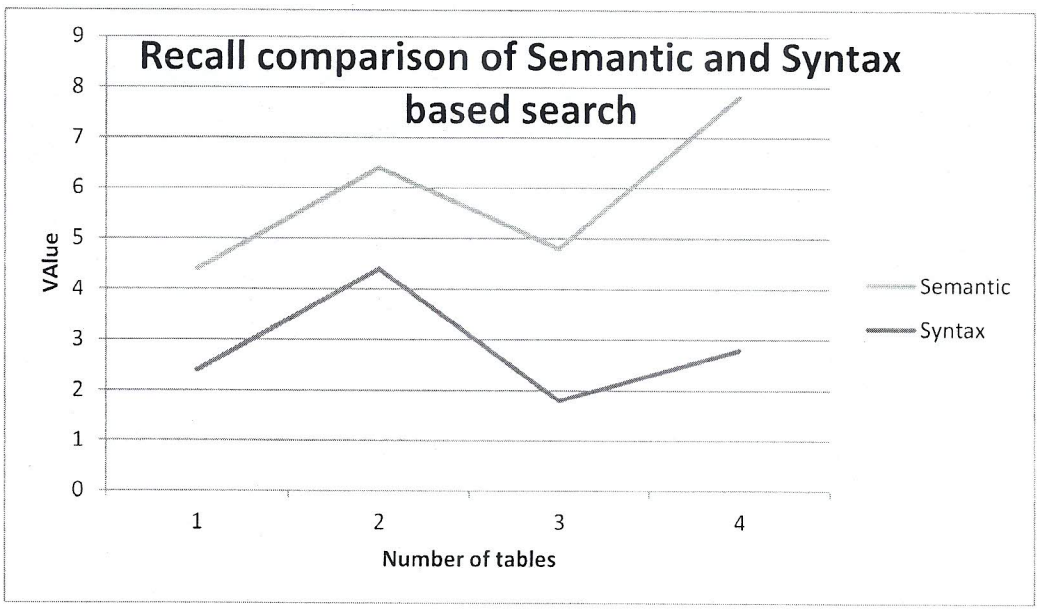

Figure 12: Recall comparison graph

Findings of the Research

The developed system was able to use Semantic knowledge to retrieve some result that captures the user intention, despite the fact that it is not categorically stated in the query. The result displayed in figure 9, 10 and 11, was possible because of the knowledge that music system, washing machine and television, are all synonymous to stereo, washer and telly respectively. The above results show an enhancement in the ability of retrieving concept based not only on exact meaning, but also on the existence of semantic relationships between the label terms.

Conclusion and Future Work

The technological advancement, among other things, has made available different databases in the domain of e-commerce. This paper present an approach for ontology-based data integration system, aimed at facilitating seamless transparent access to different databases. The relational databases were considered in their original form and format and make use of WordNet to extract potential relation between the terms. The designed application that integrate data sources and WordNet adopting the three layer architecture format. The WordNet module solves the problem of different name representations of same entity in the system, by providing terms that are synonymous to the user given query. The experimental evaluation turned out to be fairly effective in practical applications, with a particular set of data using knowledge of relations between words. The prototype designed in this paper has only been tested with simple terms, and has the capability to answer single term query. This is possibly the weakness of the developed system, as certain terms may only be relevant when seen in connection with one or more other terms. An enhancement on the system to accommodate more than a single term query will further improve the performance of the system.

References

Amshakala, K., & Nedunchezhian, R. (2011). WordNet ontology based query reformulation and optimization using disjunctive clause elimination. *International Journal of Database Management System (IJDBM)*, 3(4), 55-62

Asfari, O., Doan, B. L., Bourda, Y., & Sansonnet, J. P. (2009). Personalized access to information by query reformulation based on the state of the current task and user profile. Proceed of the third International Conference on Advances in Semantic Processing, held in Malta. 113-116.

Ghawi, R., & Cullot, N. (2009). Building ontologies from XML data sources. 1st International Workshop on modeling and visualization of XML and Semantic Web Data. Held in Linz- Austria. Pp. 480 – 484.

Grigoris A., & Frank, H. (2008). A semantic web primer, second edition. Massachusetts London: MIT press Cambridge.

Guarino, N., & Giaretta, P. (1995). *Ontologies and knowledge bases: Towards a terminological clarification towards very large knowledge bases.* Amsterdam: IOS Press. Pp. 25-32.

Halevy, A. (2001). Answering queries using views: A survey. *VLDB Journal*, 10(4),270–294.

Jayaprabha, P., & Somasundaram, R. M. (2011). Optimized semantic search through categorization of semantic web. *European Journal of Scientific Research,* 62(1), 128-141.

Khan, A. H., Minhas, A. A., & Niazi, M. F. (2008). Representation of UML activity models as ontology. Accepted in *5th International Conference on Innovations in Information Technology*, Dubai, UAE.

Maurizio, L. (2002). Data integration: A theoretical perspective. *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* Pp. 233–246.

Patwardhan, S., & Pedersen, T. (2006). Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1-8). Trento, Italy.

Raji, G. (2010). *Ontology-based cooperation of information system.* Unpublished Ph.D. thesis de Bouegogne University, France.

Seksun S., Djamal, B., Pierre-Antoine, C., & Philippe, T. (2005). Wrapping and integrating heterogeneous database with OWL. Proceed of International conference on Enterprise Information Systems: (ICEIS), Miami- U.S.A. Pp. 145-150.

Sharma, A. K., & Payal, G. (2010). Ontology driven query expansion for better image Retrieval. *International Journal of Computer Applications,* 5(10), 33-37

Vivisimo Clustering Engine (2002). http://vivisimo.com. Access 10th September, 2015.

Waris, A., Sharifullah, K. (2005). Ontology driven query expansion in data integration. Fourth International IEEE Conference on Semantics, Knowledge and Grid. Pp. 57-63.

Wordnet Documentation (2002). http://wordnet.princeton.edu. Access 5th August, 2015.