Demo of *SemIndex*: Semantic-Aware Inverted Index on Text

Joe Tekli[1], Richard Chbeir[2], Yi Luo[3], Marc Al Assad[1], Carlos Raymundo Ibanez[4], Agma J. M. Traina[5], Caetano Traina Jr.[5], Kokou Yetongnon[3]

[1] ECE Department, Lebanese American University, Byblos, Lebanon
[2] IUT of Bayonne, University of Pau and Adour Countries, Anglet, France
[3] LE2I Laboratory UMR-CNRS, University of Bourgogne, Dijon, France
[4] School of Engineering, Universidad Peruana de Ciencias Aplicadas, Lima, Peru
[5] ICMC, University of Sao Paulo, Sao Carlos-SP, Brazil

Processing keyword-based queries is a central problem in Information Retrieval (IR), where several studies have been done to develop effective keyword-based search techniques [1, 2]. A standard containment keyword-based query, which retrieves textual identities that contain a set of keywords, is generally supported by a full-text index. The inverted index is considered as one of the most useful full-text indexing techniques for large textual collections [3], supported by many RDBMSs[1]. It is also increasingly used on semi-structured [2] and unstructured data [1] to support keyword-based queries.

Besides the standard containment keyword-based query, the so-called semantic-aware or knowledge-aware (keyword) query has emerged as a natural extension, encouraged by (non-expert) user demand. In semantic-aware queries, some knowledge[2] needs to be taken into consideration while performing query processing. To illustrate this, let's assume having data from a movie database Δ, as shown in Table 1. Each movie in Δ, identified with an *id*, is described with some text, including movie *title* , *year* and *plot*. For queries *"sound of music"*, *"Maria nun"* and *"sound Maria"*, the query result is movie O_3. However, if the user wants to search for a movie but cannot recall the exact movie title, or the exact plot description, it is natural to assume that she may modify the query terms to some semantically similar terms, for example, *"voice of music"*.

Table 1. Sample Movie data collection extracted from IMBD[3].

ID	Textual content
O_1	*When a Stranger Calls (2006)*: A young high school student babysits for a very rich family. She begins to receive strange phone calls threatening the children...
O_2	*Days of Thunder* (1990): Cole Trickle is a young racer from California with years of experience in open-wheel racing winning championships in Sprint car racing…
O_3	*Sound of Music, The (1965)*: Maria had longed to be a nun since she was a young girl, yet when she became old enough discovered that it wasn't at all what she thought...

Also, it is common that the terms provided by users are not exactly the same, but are semantically similar/related to terms that the plot providers use. In addition, the movies might not be extensively described or well-tagged in the database, or might not be described using the same attributes (e.g., NoSQL database). Clearly, the standard inverted index which only supports exact term matching cannot deal with these cases. However, it would be possible to solve the above problems by answering semantic-aware queries, if semantic knowledge can be somehow combined into query processing.

Various approaches combining different types of data and semantic knowledge have been proposed to enhance query processing (briefly described below). In this study, we present a new approach called *SemIndex* integrating knowledge into an inverted index to support semantic-aware querying. Major differences between our work and existing methods include:

[1] Relational Database Management Systems
[2] Also called domain, collaborative, collective, or semantic knowledge, based on the application domain at hand.
[3] Internet Movie DataBase, available from http://www.imdb.com/

- **Pre-processing the index**: Enclosing semantic knowledge directly into an inverted index, so that the semantic-aware processing tasks can be done prior to query processing, in comparison with query rewriting/relaxation and query suggestion [4, 5], or query result organization [6, 7].
- **User involvement**: Allowing end-users to write classical queries as well as semantically enriched queries such that the users are involved in the whole process: during initial query writing, and then query rewriting, in comparison with existing works where users are only involved in the query refinement (expansion, filtering, etc.) process [8, 9].
- **Providing more relevant results**: identifying *more* semantically relevant results than what a traditional inverted index could provide, while doing it more efficiently than existing semantic disambiguation techniques [6, 10], which: i) usually require substantial processing time [11], and ii) depend on the query/data context which is not always sufficiently available [12, 13].

In this study, we explore the idea of merging a pre-existing general purpose semantic network into a standard inverted index, and design our semantic-aware inverted index, *SemIndex*, accordingly. To do so, we map into a single data structure two data resources, *a textual data collection* (represented as a traditional inverted index), and *a semantic knowledge base* (represented as a traditional semantic network). An extended query model with different levels of semantic awareness is defined, so that both semantic-aware queries and standard containment queries are processed within the same framework. Figure 1 depicts the overall framework of our approach and its main components. The *Indexer* manages *SemIndex*, while the *Query Processor* accepts semantic-aware (or standard) queries and processes the queries with *SemIndex*.

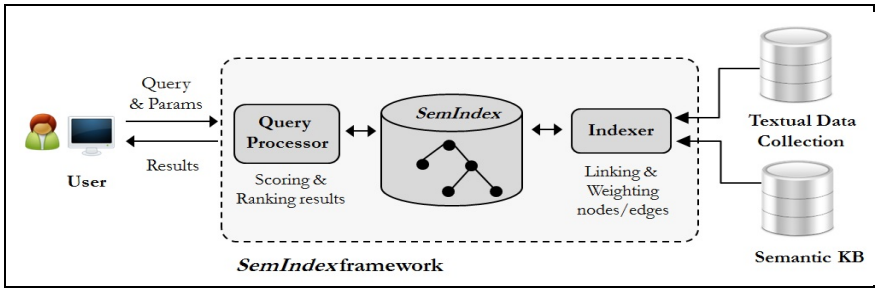

Figure 1. Overall architecture of *SemIndex* framework.

The groundwork results and overall architecture of *SemIndex* have been described in [14], and an extended experimental study has been submitted for publication in an international journal [15]. In this demonstration, we aim to showcase our systems' effectiveness and efficiency in textual DB indexing and performing semantic-aware querying, while emphasizing its logical design, and its physical implementation using a commercial RDBMS and related functionality.

References

[1] Das S., et *al.*, *Making unstructured data sparql using semantic indexing in oracle database.* IEEE ICDE Conf., 2012. pp. 1405–1416
[2] Florescu D. et *al.*, *Integrating Keyword Search into Xml Query Processing,* . Computer Networks, 2000. 33(1-6):119–135.
[3] Frakes W.B. and R.A. Baeza-Yates, *Information retrieval: Data structures and algorithms.* Prentice-Hall, 1992.
[4] Burton-Jones A. et *al.*, *A Heuristic-Based Methodology for Semantic Augmentation of User Queries on the Web.* ER, 2003, 476–489.
[5] Cimiano P. et *al.*, *Towards the Self-Annotating Web.* Inter. World Wide Web Conference (WWW'04), 2004. pp. 462-471.
[6] Navigli R. and Crisafulli G., *Inducing Word Senses to Improve Web Search Result Clustering.* EMNLP, 2010, 116–126, MIT, USA.
[7] Sinh Hoa Nguyen et *al.*, *Semantic Evaluation of Search Result Clustering Methods.* Intelligent Tools for Building a Scientific Information Platform, Studies in Computational Intelligence Volume 467, 2013. 467(393-414).
[8] Chandramouli K., et *al.*, *Query Refinement and user Relevance Feedback for contextualized image retrieval.* VIE, 2008, 453 - 458.
[9] Mishra C. and Koudas N., *Interactive Query Refinement.* Inter. Conf. on Extending Database Technology (EDBT), 2009, 862-873.
[10] Li Y. et *al.*, *Term Disambiguation in Natural Language Query for XML.* Inter. FQAS Conf., 2006. LNAI 4027, pp. 133–146.
[11] Navigli R., *Word Sense Disambiguation: a Survey.* ACM Computing Surveys, 2009. 41(2):1–69.
[12] de Lima E.F. and Pedersen J.O., *Phrase Recognition and Expansion for Short, Precision biased Queries based on a Query Log.* 22nd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, 1999. pp. 145-152, Berkeley, CA.
[13] Voorhees E.M., *Query Expansion Using Lexical-Semantic Relations.* Inter. ACM SIGIR Conf., 1994. pp. 61-69.
[14] Chbeir R., et al., *SemIndex: Semantic-Aware Inverted Index.* Inter. ADBIS Conf., 2014. pp. 290-307.
[15] Tekli J., et al., *SemIndex: Semantic-Aware Inverted Index on Textual Data.* Submitted to Springer VLDB Journal, 2015.