SEMANTIC HMC FOR BUSINESS INTELLIGENCE USING CROSS-REFERENCING

Rafael PEIXOTO
Checksem - Laboratoire Le2i, UMR CNRS 6306, Dijon, France
rafpp@isep.ipp.pt
Thomas HASSAN
Checksem - Laboratoire Le2i, UMR CNRS 6306, Dijon, France
thomas.hassan@u-bourgogne.fr
Christophe CRUZ
Checksem - Laboratoire Le2i, UMR CNRS 6306, Dijon, France
christophe.cruz@u-bourgogne.fr
Aurélie BERTAUX
Checksem - Laboratoire Le2i, UMR CNRS 6306, Dijon, France
aurelie.bertaux@u-bourgogne.fr
Nuno SILVA
GECAD, ISEP-IPP, Porto, Portugal
nps@isep.ipp.pt

Abstract. *Keeping abreast with current market trends requires the centralization of large amount of information. Due to the increasing number of news available on the web, selecting only valuable information for each consumer is mandatory to reduce the consumer information overload. However, information available on the web can have uncertain and imprecise data, leading to veracity issues. We aim to measure Big Data veracity using cross-referencing of several information sources. In this work we present a new vision to cross-referencing of several huge web information sources using a Semantic Hierarchical Multi-label Classification process called Semantic HMC to extract the knowledge available in those sources.*

Keywords: Ontology, Hierarchical Multi-label Classification, similarity measure.
JEL classification: L86 Information and Internet Services

1. Introduction

The decision-making process in the economic field requires the centralization and intakes of a large amount of information. The aim is to keep abreast with current market trends. Thus, contractors, businessmen and salespersons need to continuously be aware of the market conditions. This means to be up-to-date regarding ongoing information and projects undergoing development. With the help of economic monitoring, prospects can be easily identified, so as to establish new contracts. Our tool called First Pro'fil [1]–[3] (http://www.firsteco.fr/) is specialized in the production and distribution of press reviews about French regional economic actors. The overload of news is a particular case of information overload, which is a well-known problem, studied by Information Retrieval and Recommender Systems research fields. News recommender systems already exist [4], Athena [5], GroupLens [6] or News Dude [7]. Some of these systems use domain knowledge to improve the recommendation task [4], [5]. To achieve this goal, a content-based recommender system is being developed [3], [8]. A recommender system is necessary for ranking the items and a content-based approach is required to analyze the content of each article to structure and preserve information content. The results of the analysis enables linking the domain knowledge to the articles to improve the recommendation task [4], [5].

However the amount of news available on the web is growing, requiring new forms of processing to enable enhanced decision making, insight discovery and process optimization. The term of Big Data is mainly used to describe these huge datasets. Various types of data compose Big Data, including unstructured data that represents 90% of its content [10]. An increasing number of V's has been used to characterize Big Data [9], [10]: Volume, Velocity, Variety, Veracity and Value. Volume means the big amount of data that is generated and stored by transaction-based data stored through the years, text data constantly streaming in from social media, increasing amounts of sensor data being collected, etc. Thus, Big Data is not only a huge volume of data but it must be processed quickly. Velocity means both (i) how fast data is being produced and (ii) how fast the data must be processed and analyzed to meet a demand. Variety means that various types of data compose Big Data. These types include semi-structured and unstructured data such as audio, video, webpages, and text, traditional structured data, etc. Veracity concerns the truthfulness in data. In traditional data warehouses there was always the assumption that the data is certain, clean, precise and complete but in Big Data context, namely the user-generated data can be uncertain, erroneous, imprecise and incomplete. The Value characteristic measure how valuable is the information to a Big Data consumer. Value is the desired outcome of Big Data analytics and Big Data "raison d'etre" because if data don't have value then is useless. We aim to measure data veracity of a Big Data source by using similar data in several web sources linked by cross-referencing. Cross-referencing means linking the several textual information sources that share similar meanings. When focusing on cross-referencing web information sources, one must instantly focus on extracting knowledge from these sources.

To extract knowledge from Big Data sources we propose to use a Semantic HMC [11], [12] process that is capable of Hierarchically Multi-Classify a large Variety and Volume of unstructured data items. Hierarchical Multi-Label Classification (HMC) is the combination of Multi-Label classification and Hierarchical classification [13]. In HMC items can be assigned to different hierarchical paths and simultaneously may belong to different class labels in a same hierarchical level. The Semantic HMC process is unsupervised such that no previously labelled examples or enrichment rules to relate the data items with the labels are needed. The label hierarchy and the enrichment rules are automatically learned from the data through scalable Machine Learning techniques.

This paper claims that cross-referencing high quality information (i.e. events) in items from several sources we can measure its veracity. The main contribution of this paper is then to cross-referencing of huge web information sources by using the Semantic HMC to extract the knowledge available in these sources. The cross-references are then used to measure the data veracity and improve the recommendation of economical news.

Next section focuses on related work such as Semantic HMC, semantic measures and Cross-Referencing Methods Proposals. Section 3 proposes how to cross-referencing huge web information sources using the Semantic HMC for veracity measure. The last section concludes this paper.

2. Related work

In order to compare cross-referencing methods of web information sources, we need to evaluate the Semantic Measure between concepts. Semantic Measure is normally a philosophic term. It is a point of view that differs from one person to another regarding the semantic links strengths between two concepts. Trying to computerize this philosophic term, in order to compare different textual information, is a complex task and requires performing high-level language processing. However, the evaluation of the Semantic Measure between two concepts depends firstly on the kind of the semantic links, and secondly on the kind of knowledge resources.

2.1 Semantic Measure Type

In order to compare two concepts, and in particular two textual information sources in the case of documentary research, one must evaluate the semantic measure between these sources. Semantic measure is a generic term covering several concepts [14]:

- **Semantic relatedness**, is the most general semantic link between two concepts. Two concepts do not have to share a common meaning to be considered semantically related or close, as they can be semantically linked (related) by a functional relationship or frequent association relationship like meronym or antonym concepts. (e.g. Pilot “is related to” Airplane).
- **Semantic similarity**, is a specific case of semantic relatedness. Two concepts are considered similar if they share common meanings and characteristics, like synonym, hyponym and hypernym concepts (e.g. Old “is similar to” Ancient).
- **Semantic distance**, is the inverse of the semantic relatedness, as it indicates how much two concepts are unrelated to one another.

2.2 Cross-Referencing Methods Proposals

We have identified three kinds of approaches for Cross-Referencing information: semantic similarity, paraphrase identification and event extraction techniques.

In order to improve cross-referencing methods for web information sources, semantic measures and precise semantic similarity definitions are proposed in the literature. These measures can normally be grouped into five categories: Path Length-based measures [15], Information Content-based measures [16], [17], [18], Feature-based measures [19], [20], Distributional-based measures[21], [22] and Hybrid measures [18], [23], [24].

Paraphrase identification corresponds to the ability of identifying phrases, sentences, or longer texts that convey the same, or almost the same information [25]. Paraphrase identification techniques can be classified into three categories: recognition, generation, or extraction.

A current active research field in cross-referencing methods for web information sources is event extraction. Event extraction is a common application of text mining to derive high quality information from text by identifying events. Event extraction techniques depend on paraphrase identification methods to identify events expressed in different ways. As Hogenboom et al. [26] cite, one can distinguish between three main categories of event extraction: data-driven event extraction, knowledge-driven event extraction, and hybrid event extraction.

3. Cross-referencing

The current paper discusses a large variety of approaches to natural language processing from diverse fields. When focusing on cross-referencing web information sources, one must instantly focus on extracting knowledge from these sources. While there is a compromise between the size, the coverage, the structure and the growth of the knowledge resources, dealing with the knowledge extraction from huge web information sources is considered the main challenge in hands.

We believe that improving the web structure will be the most efficient approach to measure veracity in web information sources. We suggest to benefit from the large data offered by the web, and consider data-driven approaches as the most suitable event extraction techniques. These techniques aim at converting data into knowledge relying on quantitative methods such as clustering and the use of statistics. Therefore, choosing paraphrase extraction approaches based on distributional hypothesis and on recognition approaches based on surface string similarity, are a good solution as they both depend directly on semantic measures, which can be used to benefit from the presence of large context like the use of distributional-based measures. Since no approach is yet proved as the most efficient and reliable, one must choose, regarding the context of the issue, the most suitable combination of approaches. This choice depends in the first place on the knowledge resource used, then the event extraction technique. Finally, it depends on the best match between the paraphrase identification techniques and the similarity measure.

The proposed Semantic HMC process [11], solves this issue by learning automatically a concept hierarchy and enrichment rules from Big Data through scalable Machine Learning techniques. To represent the knowledge in the Semantic HMC process, an Ontology-described Knowledge Base is used. Ontologies [27] are the most accepted way to represent semantics in the Semantic Web [28] and a good solution for intelligent computer systems that operate close to the human concept level, bridging the gap between the human requirements and the computational requirements [29]. Initially the Semantic HMC enriches the ontology from the huge Volume and Variety of initial data and once this learning phase is finished, the classification system learns incrementally from the new incoming items to provide high Velocity learning. The result of this Semantic HMC process is a rich ontology with the items classified with the learned concept hierarchy.

To infer the most specific concepts for each data item and all subsuming concepts, rule-based reasoning is used, exhaustively applying a set of rules to a set of triples to infer conclusions. This Rule-based reasoning approach allows the parallelization and distribution of work by large clusters of inexpensive machines using Big Data technologies as Map-reduce [30].[30] Web Scale Reasoners [31] currently uses Rule-Based reasoning to reach high scalability by loading parallelization and distribution, thus addressing the Velocity and Volume dimensions of Big Data.

The Semantic HMC process consists of 5 individually scalable steps matching the requirements of Big Data processing:

- Indexation creates an index of parsed items and identifies relevant terms.
- Vectorization creates a term co-occurrence frequency matrix of all indexed items and a TF-IDF vector of each item.
- Hierarchization creates a hierarchy of relevant concepts based on term-frequency.
- Resolution creates classification rules to enrich the ontology based on term-frequency.
- Realization first populates the ontology with items and then determines the corresponding hierarchy concept and all subsumant concepts. This is intended as Hierarchical Multilabel Classificaiton (HMC).

Once the items are classified, after the realization step, a set of classified items is available for post processing. Similarity measures can then be easily computed between items classified with the same labels. Notice that while paraphrase and event extraction technics cannot be undertaken directly on the whole set of items available in one day at once, it can be undertaken with smaller subsets (i.e. classified with the same labels) and then cross-reference their sources. Using these cross-referenced items we can measure its veracity. Two uses of cross-referencing for measure item veracity are identified:

- Cross-referencing with information sources that have a particular trustworthy. As an example, if the item is cross-referenced with trusted sources it is a veracity indicator.
- Cross-referencing with a significant set of items from several sources. As an example, if the item is cross-referenced from several sources we can state that this item has a higher veracity than an event that appears in a restricted number of items from only one source.

Then exploiting the veracity of its items easily does easily measure the veracity of each source.

4. Conclusions

In this paper we present how to cross-referencing of large web information sources by using a Semantic HMC process to extract the knowledge available in these sources. This cross-referencing principle allows the First Eco Pro'fil to analyze the economical news veracity. It also discusses a large variety of approaches to natural language processing from diverse fields that are mandatory to do cross-referencing. In further work we aim to measure the data veracity as described and use it in the value extraction process. Our current work consists in the implementation of the proposed methodology using programming models for processing and generating large data sets as Map-Reduce.

Acknowledgment

This project is founded by the company Actualis SARL, the French agency ANRT, the "Conseil Régional de Bourgogne" and through the COMPETE Program under the project AAL4ALL (QREN13852).

5. References

[1] C. Cruz and C. Nicolle, "Ontology Enrichment and Automatic Population From XML Data," *Learning*, pp. 17–20, 2008.

[2] D. Werner and C. Cruz, "Precision difference management using a common sub-vector to extend the extended VSM method," *Procedia Comput. Sci.*, vol. 18, pp. 1179–1188, 2013.

[3] D. Werner, N. Silva, and C. Cruz, "Using DL-Reasoner for Hierarchical Multilabel Classification applied to Economical e-News," in *Science and Information Conference,* 2014, p. 8.

[4] D. C. De Roure, S. E. Middleton, and N. R. Shadbolt, "Ontological user profiling in recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1. pp. 54–88, 2004.

[5] W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom, "Ontology-based news recommendation," in *Proceedings of the 1st International Workshop on Data Semantics - DataSem '10*, 2010, p. 1.

[6] P. Resnick, N. Iacovou, and M. Suchak, "GroupLens: an open architecture for collaborative filtering of netnews," *Proc. ...*, vol. pp, pp. 175–186, 1994.

[7] D. Billsus, D. Billsus, M. J. Pazzani, and M. J. Pazzani, "A personal news agent that talks, learns and explains," in *Proceedings of the third annual conference on Autonomous Agents*, 1999, pp. 268–275.

[8] D. Werner, C. Cruz, and C. Nicolle, "Ontology-based Recommender System of Economic Articles.," in *WEBIST*, 2012, pp. 725–728.

[9] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, Jan. 2014.

[10] P. Hitzler and K. Janowicz, "Linked data, big data, and the 4th paradigm," *Semant. Web*, vol. 4, pp. 233–235, 2013.

[11] T. Hassan, R. Peixoto, C. Cruz, A. Bertaux, and N. Silva, “Semantic HMC for big data analysis,” in *Big Data (Big Data), 2014 IEEE International Conference on*, 2014, pp. 26–28.

[12] T. Hassan, R. Peixoto, C. Cruz, N. Silva, and A. Bertaux, “Extraction de la Valeur des données du Big Data par classification multi-label hiérarchique sémantique,” in *EGC 2015 - 15ème conférence internationale sur l’extraction et la gestion des connaissances*, 2015.

[13] W. Bi and J. Kwok, “Multi-label classification on tree-and DAG-structured hierarchies,” *Yeast*, pp. 1–8, 2011.

[14] A. Budanitsky and G. Hirst, “Evaluating WordNet-based Measures of Lexical Semantic Relatedness,” *Comput. Linguist.*, vol. 32, no. August 2005, pp. 13–47, 2006.

[15] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification,” *An Electron. Lex. Database*, pp. 265–283, 1998.

[16] D. Lin, “An Information-Theoretic Definition of Similarity,” in *Proceedings of ICML*, 1998, pp. 296–304.

[17] P. Resnik, “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language,” *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, 1999.

[18] T. Pedersen and J. Michelizzi, “WordNet :: Similarity - Measuring the Relatedness of Concepts,” *HLT-NAACL--Demonstrations ’04 Demonstr. Pap. HLT-NAACL 2004*, no. Patwardhan 2003, pp. 38–41, 1998.

[19] A. Tversky, “Features of similarity.,” *Psychological Review*, vol. 84. pp. 327–352, 1977.

[20] E. G. M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, “X-Similarity: Computing Semantic Similarity between concepts from different ontologies,” *J. Digit. Inf. Manag.*, vol. 4, pp. 233–237, 2006.

[21] R. L. Cilibrasi and P. M. B. Vitanyi, “The Google Similarity Distance,” *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, 2007.

[22] D. Hindle, “Noun Classification from predicate-argument structures,” in *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, 1990, pp. 268–275.

[23] R. Knappe, H. Bulskov, and T. Andreasen, “On Similarity Measures for Concept-based Querying,” in *Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA’03)*, 2003, pp. 400–403.

[24] Z. Zhou, Y. Wang, and J. Gu, “New model of semantic similarity measuring in wordnet,” in *Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering*, 2008, pp. 256–261.

[25] I. Androutsopoulos and P. Malakasiotis, “A survey of paraphrasing and textual entailment methods,” *J. Artif. Intell. Res.*, vol. 38, pp. 135–187, 2010.

[26] A. Syed, K. Gillela, and C. Venugopal, “The Future Revolution on Big Data,” *Future*, vol. 2, no. 6, pp. 2446–2451, 2013.

[27] T. R. Gruber, “A Translation Approach to Portable Ontology Specifications by A Translation Approach to Portable Ontology Specifications,” *Knowl. Creat. Diffus. Util.*, vol. 5, no. April, pp. 199–220, 1993.

[28] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001.

[29] L. Obrst, “Ontologies for semantically interoperable systems,” in *Proceedings of the twelfth international conference on Information and knowledge management - CIKM ’03*, 2003, pp. 366–369.

[30] J. Dean and S. Ghemawat, “MapReduce : Simplified Data Processing on Large Clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 1–13, 2008.

[31] J. Urbani, “Three Laws Learned from Web-scale Reasoning,” in *2013 AAAI Fall Symposium Series*, 2013.