AN ONTOLOGY-BASED RECOMMENDER SYSTEM FOR ECONOMICAL E-NEWS

David WERNER
Université de Bourgogne, LE2I, CNRS
david.werner@u-bourgogne.fr
Nuno SILVA
GECAD and School of Engineering Polytechnic of Porto
nps@isep.ipp.pt
Christophe CRUZ
Université de Bourgogne, LE2I, CNRS
christophe.cruz@u-bourgogne.fr
Aurélie BERTAUX
Université de Bourgogne, LE2I, CNRS
aurelie.bertaux@u-bourgogne.fr

Abstract. *This paper focuses on a recommender system of economic news articles. Its objectives are threefold: (i) automatically multi-classify new economic articles, (ii) recommend articles by comparing profiles of users and multi-classification of articles, and (iii) managing the vocabulary of the economic news domain to improve the system based on seamlessly intervention of documentalists. In this paper we focus on the automatic multi-classification of the articles, managed by inference process of ontologies, and the enrichment of the documentalist-oriented ontology which provides the necessary capabilities to the DL reasoner for automatic multi-classification.*

Keywords: multi-classify, recommender system, ontology economical e-news, machine learning.
JEL classification: Innovation and Invention: Processes and Incentives O31

1. Introduction

The decision-making process in the economic field requires centralization and intake 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 concerning information and development of projects. With the help of economic monitoring, prospects can be easily identified, so as to establish new contracts. Our tool is specialized in the production and distribution of press reviews about French regional economic actors (fig. 1).

The overload of news information 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 [1]: Athena [2] GroupLens [3] or News Dude [4]. Some of these systems use domain knowledge to improve the recommendation task [2], [1]. A recommender system is necessary for the item ranking 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 enable to link the domain knowledge to the articles to improve the recommendation task [2], [1]. Content-based recommender systems typically follow a two-step process: (i) the indexing of articles (also known as *classification*)

and users (also known as *profiling*), and (ii) the comparison process which consists in comparing classification and profiling. The latter computes the article relevance with regards to the user profile.
In order to capture this economical context, we are moving towards a customized review for each user and towards an opinion survey on magazine readers that cover a broad array of subjects, including news services. Criteria for a relevant customization of the review were identified as a result of this survey as well as expert domain knowledge. These criteria are economic themes (i.e. main economic events), economic sectors, temporal and localization information. Our domain knowledge is based on expert-defined thesaurus and an ontology allowing managing the main concepts and relations. As consequence of this effort the complete production process of the review was redesigned to produce and to automatically distribute a customized review for each user. So, the aim of the overall system is to manage all news articles produced, and provide the most relevant article for each customer. Yet, while the comparison process is automated, the article classification process remains manual. While this is a time consuming process, it ensures the quality of the recommendation as long as the documentalists know the economical context and perform the classification. The aim of this paper is to show our last results by proposing an automatic classification process that mimics process of the documentalists.
The paper is organized as follows. First, we present the work context of recommender systems and the overall system proposal. The third section presents the hierarchical multi-label classification and the application of this principle in our four step method. The fourth section presents some evaluations. Finally, we summarize the contributions.

2. What is Context?

A recommender system aims at providing for each user the better items according to his/her needs. Items can be websites, news articles, books, video, music, washing machine, etc. In the recommender system literature two paradigms are distinguished. First, content-based recommender systems (CB) try to recommend items similar to those a given user has liked in the past without the feedback of other users. This specificity is required for news recommender systems. Second, collaborative filtering recommender systems (CF) identify users whose preferences are similar to those of the given user and recommend items they have liked [2]. Some subtasks should be performed and the first is named *indexing process*. It is possible to distinguish two cases, the indexing of items, by content analysis, and the indexing of profiles, via profiles learning which generally includes a study of the behavior for implicit feedbacks or proposed to the user to giving explicit feedbacks as “I like” button. Both can be seen as a multi-classification task. The second process, *comparison*, consists in filtering each item relative to a given profile [3]. Usually, both types of recommender systems are combined to define a so-called Hybrid recommender systems and to overcome drawbacks of each case. The survey of K. Nagewara Rao [4] proposed a general comparison of the main advantages and drawbacks of each kind of recommender systems. In this paper we focus on content-based recommender system, and on the indexing subprocess of textual items, based on a controlled vocabulary (e.g. fig. 1 red square).

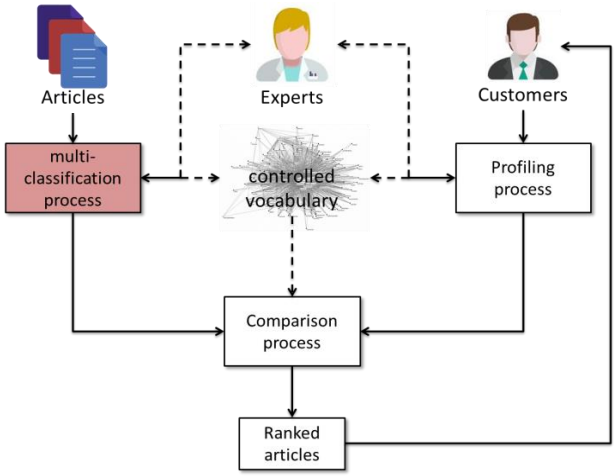

Figure 1. Our ontology-based recommender system

The indexing process (or multi-classification process e.g. fig. 1) consists in selecting a set of keywords from the controlled vocabulary and associating it with an item (content analysis) or a user profile (profiling). As it is presented in the survey about controlled vocabularies from [5], more and more companies use controlled vocabularies in their information system. Different kinds of structure are used to manage vocabularies, i.e. from the lowest to the richest semantic definition: glossary, taxonomy, thesaurus, ontology. A lot of companies plan to use ontologies in their applications [5]. While acquiring, managing and maintaining controlled vocabularies are important yet relatively easy tasks for the documentalists, the ontology approach to model the domain knowledge is hardly accessible to them due to the complexity of the logical structure. Actually, this model differs from their habits, which is much more related to a taxonomy (subsumption relations between entities) or a thesaurus (different relations between entities, notably words). Through the remaining of the text we will use the terms "documentalist" and "expert" interchangeably.
Our thesaurus (controlled vacabulary e.g. fig. 1) has been captured in a very light, expert-oriented fashion, with minimal formal semantics and consistency obligations. For that, the experts have been provided with only some types of relation as defined in SKOS [6]. On the contrary, for the authoring of this ontology the strict semantics of OWL DL has been followed. Each of the resulting thesaurus was appended to the Theme and Sector classes, by interpreting (i) each label as a class, and (ii) the "skos:broader" relationships as owl:subClassOf relationships, thus forming a taxonomy (hierarchy of subsumption relationships) of labels (i.e. a taxonomic thesaurus).
The comparison of the item classification with the user profile is performed using the classification of both according to the thesaurus (fig. 1). More details about the comparison process can be found in the [7].

3. Hierarchical multi-label classification

In order to make the classification automatic, the system has to associate a set of labels from the taxonomic thesaurus to each article. This cannot be done without the two following process: (i) a text analysis process to extract keywords and other features from texts and (ii) a machine learning process to learn the classification process from examples.

i). Feature extraction processes range from simple term extraction process like tf-idf [8] to text-based semantic-aware processes, e.g. term extraction from text-based on information retrieval methods [9], or based on NLP works [10] [14]. It is possible to use different degrees of text processing tools (and preprocessing), to extract noun

phrase (i.e. tokenizer, part of speech tagger, handmade patterns and even parsers). This process allows to interlink a term set from the article-term extraction process (content analysis) and the set of taxonomic keywords for the articles indexing task.

ii). Regarding machine learning process, two main categories of label-classification prediction can be enumerated: the single-label and the multi-label classification. Single-label classification aims to learn a prediction model from a set of examples that are related with a single label from a set of disjoint labels. In multi-label classification instead, the examples are associated with a set of labels [12]. Multi-label classification faced increased attention in the last decade, overcoming the flat-label classification previous dominance, but it was only much recently that hierarchical multi-label classification (HMC) approaches received the desired attention [13], [14] [15], [16].

In [17] the authors are concerned with automatically create an ontology from the text documents without any prior knowledge about their content. For that they use an iterative and interactive 4-phases process. Unlike [17], [18] that construct thesaurus from the learning examples, in our project the thesaurus-based taxonomy already exists and should be applied in the automatic classification. Consequently, we do not aim at improving the state of the art in multi-classification, nor in ontology learning from text, but instead to propose a different method to enrich the ontology with logical axioms to multi-classify articles. As a consequence, the gap between the perspective of the expert and the classification rules representation is reduced. Our method is based on four following steps.

1. The **vectorization** phase allows generating the matrix of term frequencies from a learning set.
2. The **resolution** allows creating logical constraints (rules) associated to the keywords of the taxonomy (controlled vocabulary) using named entities extracted in the vectorization phase. This phase generates a flat ontology.
3. The **hierarchization** allows to generate a class hierarchy of subsumption of the ontology used to label documents.
4. The **realization** allows searching and deducing the most specific classes of documents to be classified which consists in generating the multi-classification.

Phases 3 and 4 are done using standard reasoner such as FaCT++, HermiT, Pellet.

4. First evaluation of the approach

In this section we present a preliminary evaluation of the approach. Due to the lack of real data for our platform, the first evaluation is based on the delicious dataset available on the Mulan project web site and already used in some multilabel-classification works [13]. It was extracted from the del.icio.us social bookmarking site on the 1st of April 2007 and contains textual features and tags of webpages. This dataset is used to train a classifier for tag recommendation.

dataset	Examples		Attributes			Label	Label
	Train	Test	Numeric	Discrete	Labels	Cardinality	Density
delicious	12920	3185	0	500	983	19.020	0.019

Table1: The dataset in some numbers

With this dataset the (**phase 1**) manual multi-classification and the (**phase 2**) feature extraction tasks are not necessary, features and tags are already associated with documents and a sub-dataset is predefined for the (**phase 3**) learning of the prediction model. The ontology is populated (**phase 4**), and some reasoners are used to perform the multi-classification task. During our evaluations different reasoners are used on different hardware

(table 2). The results produced by the reasoner are not only a multilabel-classification of documents, but also a hierarchical reorganization of tags based on the equivalence rules.

Simple rules	**FaCT++**	**HermiT**	**Pellet**
i7 4go DDR3	50s	n. e. m.[1]	n. e. m.
Xeon E3 24go DDR3	-	8h	18h
Complex rules	**FaCT++**	**HermiT**	**Pellet**
i7 4go DDR3	n. e. m.	n. e. m.	n. e. m.
Xeon E3 24go DDR3	n. e. m.	out[2]	out
Xeon E5 128go DDR3	2h/out[3]	out	out

Table 2: Reasoner time computation comparison for the ontology populated with complex rules

This table shows that the second type of rules is much more time and memory consuming. We have only one result to show. This result was produced by FaCT++, with the best machine and an ontology without any document. In 2 hours the reasoner infers a hierarchical reorganization of tags based on the equivalence rules. But the ontology populated with documents and equivalent class rules seems very time consuming even for FaCT++. The ontology with Complex rules is not evaluated in the following steps due to the lack of results provided by reasoners.

Evaluation	**Precision**	**Recall**	**F1-measure**
Proposal with Simple rules	0.3	0.6	0.18
HOMER [19]	-	-	0.25

Table 3: Evaluation and comparison with a similar work of multi-classification

Results are low, but, our rules are very naive. Another approach [16] with this dataset also shows low value for the F-measure. We can probably improve the result with more powerful rules. Simple rules are a small step to gain intelligence, but the impact on the computation time and memory used is very important.

5. Discussion and Conclusion

This paper describes the process of using an HMC approach to enrich an already existing ontology to be used for automatic multi-classification of economic news articles. We decided to capture the prediction model into the taxonomic thesaurus part of the ontology, thus transforming it into a more semantically rich ontology. Based on the early experiments, it was observed that the logical axioms/rules suggested the existence of several subsumption relations that were not present in the taxonomic thesaurus, giving rise to Direct Acyclic Graphs, i.e. a class can have more than one super-class. While this observation is potentially relevant for the refinement of the taxonomic thesaurus and therefore for the classification, a deeper and finer analysis and expert-based experiments have to be performed to better understand the advantages, disadvantages and potential applications.

Acknowledgment

This project is founded by the company Actualis SARL and the French agency ANRT.

References

[1] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, ”Ontological user profiling in recommender systems”, Acm Trans Inf Syst, vol. 22, no 1, p. 54–88, janv. 2004.

[1] Not enough memory
[2] Too much time consumption (more than 3 days)
[3] Only the hierarchical reorganization of tags for the document less ontology

[2] W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom, “Ontology-based news recommendation”, in Proceedings of the 2010 EDBT/ICDT Workshops, New York, NY, USA, 2010, p. 16:1–16:6.
[3] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: an open architecture for collaborative filtering of netnews”, in Proceedings of the 1994 ACM conference on Computer supported cooperative work, New York, NY, USA, 1994, p. 175–186.
[4] D. Billsus and M. J. Pazzani, “A personal news agent that talks, learns and explains”, in Proceedings of the third annual conference on Autonomous Agents, New York, NY, USA, 1999, p. 268–275.
[5] M. Balabanović and Y. Shoham, “Fab: content-based, collaborative recommendation”, Commun Acm, vol. 40, no 3, p. 66–72, mars 1997.
[6] P. Lops, M. de Gemmis, and G. Semeraro, “Content-based recommender systems: State of the art and trends”, in Recommender Systems Handbook, Springer, 2011, p. 73–105.
[7] K. Rao and V. Talwar, “Application Domain and Functional Classification of Recommender Systems—A Survey”, Desidoc J. Libr. Inf. Technol., vol. 28, no 3, p. 17–35, 2008.
[8] F. Kondert, T. Schandl, and A. Blumauer, “Do controlled vocabularies matter? Surevey results”, GmbH, Vienna, 2011.
[9] “SKOS Simple Knowledge Organization System Primer”, 2009. [Online]. Available: http://www.w3.org/TR/skos-primer/. [Accessed: 04-feb-2014].
[10] D. Werner and C. Cruz, “Precision difference management using a common sub-vector to extend the extended VSM method”, presented at the ICCS, International Conference on Computational Science 2013, Barcelona, Spain, 2013, p. in press.
[11] P. Cimiano, “Ontology Learning And Population from Text: Algorithms, Evaluation And Applications”. Springer-Verlag New York Inc., 2006.
[12] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval”, Inf. Process. Manag., vol. 24, no 5, p. 513-523, 1988.
[13] K. Frantzi, S. Ananiadou, and H. Mima, “Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method”. ECDL '98 Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, pages 585-604.
[14] P. Pantel and D. Lin, “A Statistical Corpus-Based Term Extractor”, in Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, London, UK, UK, 2001, p. 36–46.
[15] G. Tsoumakas and I. Katakis, “Multi-label classification: An overview”, Int J Data Warehous. Min., vol. 2007, p. 1–13, 2007.
[16] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Effective and efficient multilabel classification in domains with large number of labels”, in Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD’08), 2008, p. 30–44.
[17] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, “Decision trees for hierarchical multi-label classification”, Mach. Learn., vol. 73, no 2, p. 185-214, nov. 2008.
[18] N. Holden and A. A. Freitas, “Hierarchical classification of G-protein-coupled receptors with a PSO/ACO algorithm”, Proceedings of the IEEE Swarm Intelligence Symposium (SIS ’06), 2006.

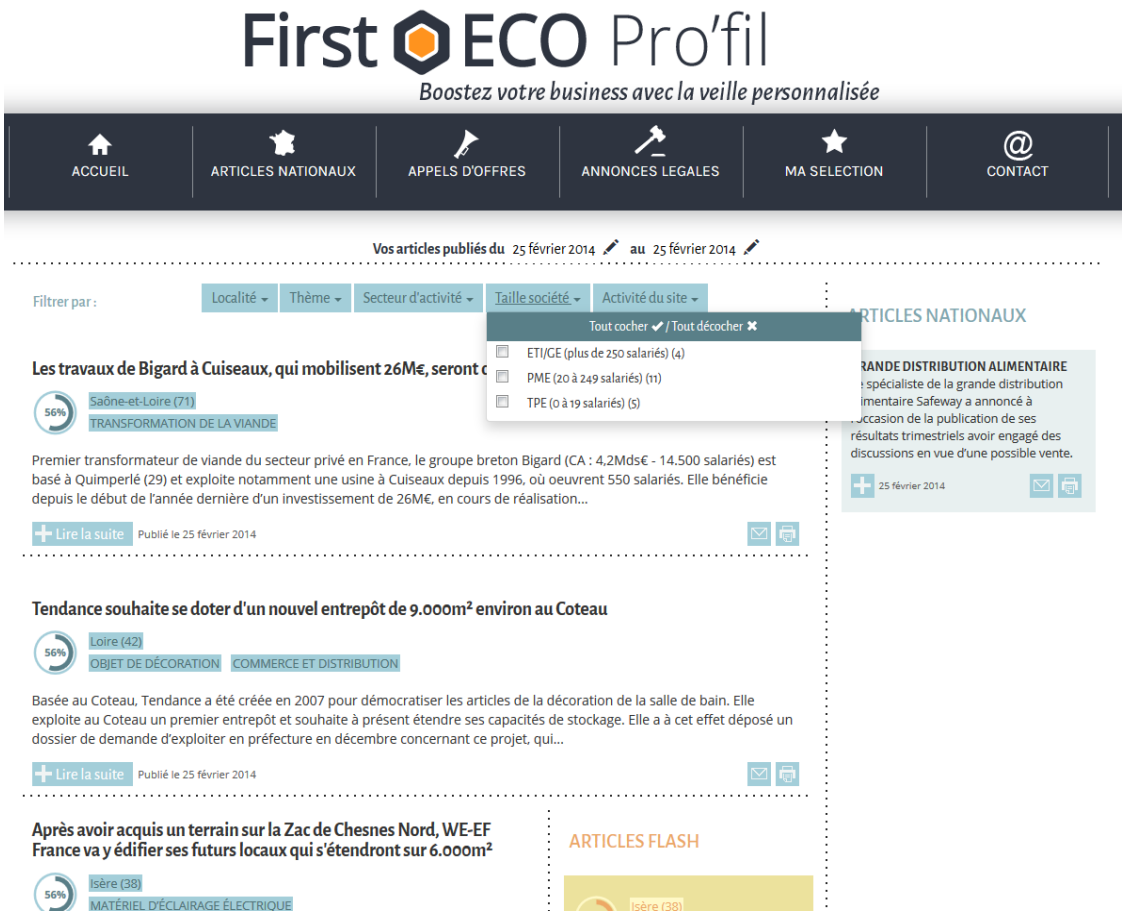

Figure 2. A snapshot of the customized e-news reviews