Information Sciences 374 (2016) 15–31

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Concept-based item representations for a cross-lingual content-based recommendation process

Fedelucio Narducci*, Pierpaolo Basile, Cataldo Musto, Pasquale Lops, Annalina Caputo, Marco de Gemmis, Leo Iaquinta, Giovanni Semeraro

Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, I-70125 Bari, Italy

ARTICLE INFO

Article history:
Received 29 January 2016
Revised 31 July 2016
Accepted 7 September 2016
Available online 12 September 2016

Keywords:
Content-based recommender systems
Concept-based representations
Wikipedia
BabelNet

ABSTRACT

The growth of the Web is the most influential factor that contributes to the increasing importance of text retrieval and filtering systems. On one hand, the Web is becoming more and more multilingual, and on the other hand users themselves are becoming increasingly polyglot. In this context, platforms for intelligent information access as search engines or recommender systems need to evolve to deal with this increasing amount of multilingual information. This paper proposes a content-based recommender system able to generate cross-lingual recommendations. The idea is to exploit user preferences learned in a given language, to suggest item in another language. The main intuition behind the work is that, differently from keywords which are inherently language dependent, concepts are stable across different languages, allowing to deal with multilingual and cross-lingual scenarios. We propose four knowledge-based strategies to build concept-based representation of items, by relying on the knowledge contained in two knowledge sources, i.e. Wikipedia and BabelNet. We learn user profiles by leveraging the different concept-based representations, in order to define a cross-lingual recommendation process. The empirical evaluation carried out on two state of the art datasets, DBbook and Movielens, shows that concept-based approaches are suitable to provide cross-lingual recommendations, even though there is not a clear advantage of using one of the different proposed representations. However, it emerges that most of the times the approaches based on BabelNet outperform those based on Wikipedia, which clearly shows the advantage of using a native multilingual knowledge source.

1. Introduction

In 1998, 70% of the content on the Web was in English [34]. Nowadays about 45% of the websites provides content in a language different from English and the number of non-English pages is rapidly growing.[1] In the past, multilingual websites were in a small number due to the high costs of development and maintenance. Companies could hardly afford those costs also because the number of non-English Internet users was really small and the potential revenues did not

* Corresponding author.
E-mail addresses: fedelucio.narducci@uniba.it, narducci@di.uniba.it (F. Narducci), pierpaolo.basile@uniba.it (P. Basile), cataldo.musto@uniba.it (C. Musto), pasquale.lops@uniba.it (P. Lops), annalina.caputo@uniba.it (A. Caputo), marco.degemmis@uniba.it (M. de Gemmis), leo.iaquinta@uniba.it (L. Iaquinta), giovanni.semeraro@uniba.it (G. Semeraro).

[1] w3techs.com/technologies/overview/content_language/all.

http://dx.doi.org/10.1016/j.ins.2016.09.022
0020-0255/© 2016 Elsevier Inc. All rights reserved.

justify the required investments [49]. However, the rapid growing of non-English Internet users is changing that scenario. In a recent statistics updated on June 30, 2015, users with the largest growth of the Internet use in the period from 2000 to 2015 are Arabic speakers (+6,091%), Russian speakers (+3,227%), Chinese speakers (+2,080%), whereas English speakers (+505%), Germans speakers (+204%), and Japanese speakers (+144%) occupy the last positions.[2] Accordingly, we can state that the Web is becoming more and more multilingual, with the top websites, such as Bing, Google, Wikipedia, etc., offering their content in hundreds of languages.

Another relevant aspect is that users themselves are becoming increasingly polyglot, i.e. people are increasingly proficient in more than one language [48]. It has been estimated that more than half of the world population is bilingual, while statistics about language education in the European Union (in 2012) show that on average 94.5% of secondary education pupils now learn English in general programs, and 50.6% learn two or more languages.[3]

According to this scenario, platforms for intelligent information access as *search engines* or *recommender systems* need to evolve in order to effectively deal with this increasing amount of multilingual information. Indeed, information retrieval (IR) systems may allow to retrieve relevant results in a language different from that used to issue the query, while information filtering (IF) systems may suggest interesting items in a language different from that the user explicitly used to express her interests. This problem is known in the literature as Cross-lingual Information Access.

This clearly motivates the need for efficient and effective IF and IR techniques that cross the boundaries of languages. In that context, we must face with the so-called *vocabulary mismatch* problem [50], i.e. relevant documents might potentially be judged as irrelevant due to a low textual overlap between query and document, or interesting items might be judged not interesting due to the low overlap between the user profile and the item descriptions. An extreme case of the vocabulary mismatch problem arises in settings where relevant (interesting) documents are written in other languages than the one of the query (user profile) [47]. One way to overcome the language barrier is to focus on the *concepts associated to words*, i.e. their *meaning*. The meaning of words is inherently multilingual, since concepts remain the same across different languages, while words used to describe those concepts in each specific language change. A concept-based representation of items and user profiles could represent an effective way to have a language-independent representation, which could act as a *bridge* among different languages.

In this paper, we investigate whether a concept-based representation is an effective strategy to provide language-independent representations of items and user profiles, which in turn allows an effective cross-lingual content-based recommendation process.

In this paper we aim at answering to the following research questions:

- **R1**: Are knowledge-based strategies able to face the content-based cross-lingual recommendation problem?
- **R2**: Are concept-based representations able to provide effective content-based cross-lingual recommendations compared to translation-based approaches?
- **R3**: Are knowledge-based representations effective to provide content-based cross-lingual recommendations when limited textual content is available?

To answer to these questions, we have performed an in-depth experimental evaluation on two state of the art datasets, i.e. `DBbook` and `MovieLens`, in order to assess the effectiveness of the cross-lingual recommendations by taking into account concept-based representations obtained by leveraging two different knowledge sources, i.e. Wikipedia and BabelNet [38], different languages, and item descriptions of different length. The results show that concept-based approaches which abstract from surface representations are suitable for cross-lingual scenarios. A clear advantage of using one of the proposed approaches did not emerge, although the use of a native multilingual knowledge source such as BabelNet often leads to better results with respect to the use of Wikipedia. Furthermore, processing shorter item descriptions leads to better results as well.

The article is organized as follows. Section 2 discusses the related work, while Section 3 describes the cross-lingual content based recommendation process. The adopted knowledge-based strategies to build language independent concept-based representations are described in Section 4. Finally, experimental results are shown in Section 5, and the conclusions are drawn in Section 6.

2. Related work

Multilingual Information Access (MLIA) and Cross-Lingual Information Access (CLIA) are the most relevant tasks for the research presented in this work. MLIA is defined as the problem of accessing, querying and retrieving information from collections in any language and at any level of specificity [42]. MLIA incorporates CLIA, which refers to technologies used for accessing a data collection in a target language l_2, by using a source language l_1, where $l_1 \neq l_2$. MLIA and CLIA have been widely investigated in the literature, in particular in the Ontology Matching and IR research areas. To the best of our knowledge, the topic of Cross-Lingual and Multilingual Information Filtering has not been properly investigated in the

[2] www.internetworldstats.com/stats7.htm.

[3] ec.europa.eu/eurostat/statistics-explained/index.php/Foreign_language_learning_statistics.