KGNR: A Knowledge-based Geographical News Recommender

Angel Luis Garrido, María G. Buey, Sergio Ilarri
IIS Department
University of Zaragoza
Zaragoza, Spain
Email: {garrido,mgbuey,silarri}@unizar.es

Igor Fũrstner, Livia Szedmina
Subotica Tech
College of Applied Studies
Subotica, Serbia
Email: {ifurst,slivia}@vts.su.ac.rs

***Abstract*—Online news reading services, such as Google News and Yahoo! News, have become very popular since the Internet provides fast access to news articles from various sources around the world. A key issue of these services is to help users to find interesting articles that match their preferences as much as possible. This is the problem of *personalized news recommendation*. Recently, personalized news recommendation has become a promising research direction and a variety of techniques have been proposed to tackle it, including content-based systems, collaborative filtering systems and hybrid versions of these two. In addition, the widespread use of mobile phones today and the different features that these phones offer users allow the possibility to keep users up to date with the latest news that have taken place in their environment, anywhere and at any time.**

This paper presents KGNR (Knowledge-based Geographical News Recommender), a new approach to develop a personalized news recommendation system as an application for mobile phones that takes into account the geolocation of the user and uses learned user profiles to generate personalized news recommendations. For this purpose, a content-based recommendation mechanism have been combined with topic-maps and geolocation for modeling the recommendation system.

I. Introduction

Nowadays, with the wide use of the Internet, more and more people prefer to read news online instead of reading the paper-format press releases. A challenging problem is how to efficiently select specific news articles from a large corpus of newly-published press releases to recommend to individual readers, where the selected news items should match the reading preference of each person as much as possible.

News aggregation websites, like Google News and Yahoo! News, collect news from different sources and provide an aggregate view of news from around the world [1]. However, a great amount of news events might be released at a rate of hundreds, even thousands, per hour. This is a critical problem with news service websites because the volumes of articles can overwhelm the users. So, the challenge is to help users find news articles that are relevant for the users and without forgetting the current news trend. Therefore, the users will receive news that matches their interests without missing the important news events, even when those events do not strictly match their particular interests. It is also important to take into account that the interests of the users may change over time, and a personalized news recommendation system should be able to incrementally update the profile of the users to reflect changes in their interests.

From the methodological perspective, the methods used in existing news recommendation systems can be categorized into three different groups: content-based methods, collaborative filtering systems and hybrid approaches [2]. In general, content-based recommendation is a methodology in response to the challenge of information overload: based on a profile of user interests and preferences, they recommend items that may be of interest to the user. This technique has been applied in various domains, such as in the context of emails, news, and web search. In the domain of news, it particularly aims at aggregating news articles according to user interests and creating a "personal newspaper" for each user. However, many content-based recommendation systems use straightforward retrieval methods such as simple keyword matching or the vector space model with basic `tf-idf` (Term frequency - Inverse document frequency) content weighting [3]. Other studies on recommendation systems have predominantly focused on collaborative filtering techniques, where user-item preference data from a group of users are used to make individual predictions. Collaborative filtering is useful in situations where content analysis is hard, such as in multimedia recommendation [4].

In addition, social networking systems, like Facebook or Twitter, and news aggregators, like Yahoo! News or Google News, are the most popular web services nowadays, which share the news feed functionality, where users of social networks and news aggregators receive a set of news from their friends and favorite news sources, respectively. The spatial situation of the user usually is covered, but many aspects could be improved,or, at least, they could receive an most appropriate treatment in order to avoid users miss important news that are spatially related to them. This issue could be relevant for complementing the functionality of this kind of software. Therefore, a location-aware news feed system enables mobile users to share geo-tagged user-generated messages, e.g., users can receive nearby messages that are the most relevant to them. For example, [5] enables users to post messages with spatial extent rather than static point locations, and takes into account their locations when computing news feeds for them.

In general, geolocation is a useful process commonly used in many information systems, but yet rarely built in Content Management Systems (CMS). In a CMS it is typical to have indexing processes and search tools which can help to quickly find the documents that contain a particular word, but this kind of software has some difficulties to face specific challenges when working with geographical issues. Therefore, integrating geolocation processes into a CMS can increase its capabilities of information management and enhance its usability from a documentary point of view. The process of geotagging aims at extracting the places or locations where a text is framed. The complexity of this type of tagging depends on the tagging level to reach, the geography that must be covered, and the context of the text.

This work focuses on content-based recommendation, by approaching the problem from an information retrieval perspective in which recommendations are based on matching textual content. The general idea is shown in Figure 1. The developed recommendation engine is designed to identify the location that is mentioned in the textual content. It is able to detect whether a word is really a location or not in the case of homonyms, and it also establishes the importance of each place/location related to the text for subsequent labeling. The system can also benefit from the context of the text to optimize its results. All of this is achieved through an open implementation which allows, on the one hand, adapting the geographical database to the context and, on the other hand, to fit the standard tagging system used. The development of KGNR is being supported by the Heraldo Group[1], a leading Spanish media, which has facilitated us technical support and contents to carry out the system development and the testing. So, the main innovation of this work is the approach to recommendation systems by applying a elaborated geolocation process, additionally combined with NLP techniques in order to extract knowledge from the news. To represent this knowledge a simple form of ontology is used: topic maps [6].

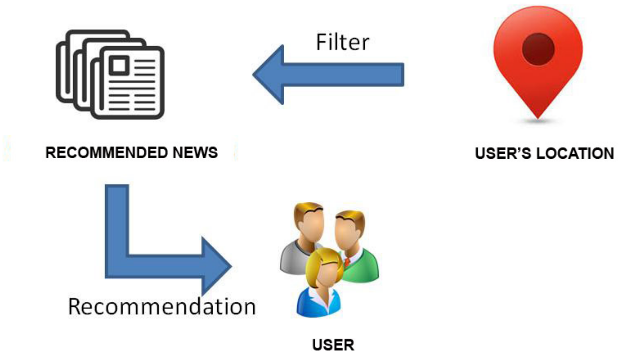

Fig. 1. Overview of the KGNR approximation

This paper is structured as follows. Section II explains the general architecture of our solution, and section III describes a case study of the application of our system. Finally, Section IV provides our conclusions and some lines of future work.

II. Architecture

The system consists of a number of working modules located on a server that receive a news flow. These services are responsible for processing and categorizing the news before distributing them to the different users. The mobile application is just a client that allows users to receive news both on demand (with a search engine that allows free searches and searches by topic) or directly using the recommendation module. News are processed on the server to obtain not only a detailed set of tags of each of them, but also a representation of its more relevant content as a knowledge graph.

Topic maps [6] have been chosen as the way of representing the embedded knowledge of a piece of news. A topic map is a simple format of knowledge-based representation supported by entities and relationships. The concepts appear in a hierarchical graph with the most inclusive and most relevant concepts at the top of the map, and the more specific and less relevant concepts at the bottom. Topic maps facilitate sense-making and learning activities, and they are used in many different fields [7].

KGNR gives recommended news to users according to several factors. First, when users register on the system, they have to select one or more predefined categories in which they are interested: politics, sports, science, cinema, music, etc. With this feature, the typical problem of initialization of recommenders known as "cold start" is avoided. Second, as the system allows for free searches, it stores in the user profile the keywords that they use to do the searches. Third, KGNR takes also into account the news that the users select and read. Then, this information can be altered by users, because the system provides the possibility of evaluating the news that they have read ("I like" / "I do not like"). KGNR takes these factors and generates dynamically, as is explained below, a profile for each user, also in a topic map format.

General categorization module: The details of this implementation can be found in [8]. Labeling techniques for obtaining common keywords by lemmatization and well statistical methods such as *tf-idf* [9] are used in the news categorization task. Moreover, KGNR allows creating a custom thesaurus in order to classify the news on general topics. The system can be trained using SVM (Support Vector Machine) [10] to automatically categorize the incoming news. In case of requiring more detailed categorizations, KGNR also incorporates a rule-based system that allows to provide a much more detailed categorization. This mechanism is supported by natural language processing tools, ontologies [11] that store the specific vocabulary, and rules of inference related to the topics to be categorized. This categorization system is dynamic and can be updated with new topics over time. This methodology has been used in real scenarios as described in [12].

Geographical categorization module: On the other hand, given the importance of the geographical location in the labeling of news, the system incorporates a powerful geographic categorization system. The aim of this module is to extract from the news a list of proper nouns that correspond to a geographical location. For that purpose, news is analyzed using a natural language processing tool, and one of the tasks performed by this software precisely to identify named entities, that may correspond to names of people, places, organizations, etc. Those named entities belonging to a gazetteer are candidates to be identified as places that must be labeled in the text.

Our work is based own well-established technologies as described in [13]. The main idea is to find the most representative toponyms using a NER (Named Entity Recognition)

[1] http://www.heraldo.es

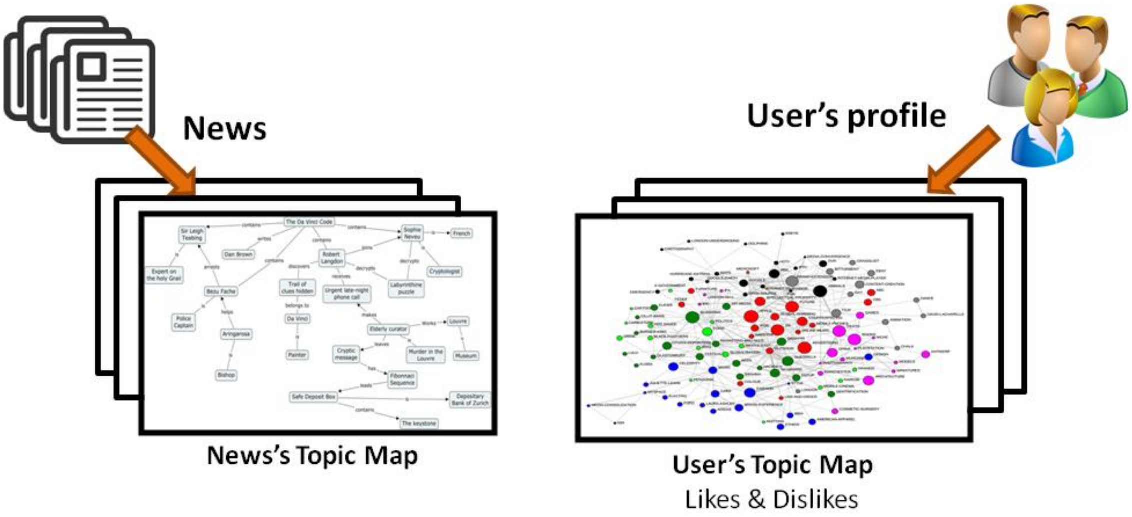

Fig. 2. The topic maps are generated from news and from users

methodology [14]. Usually this process is performed by using a *gazetteer* [15]. There are many gazetteers that can be used for this purpose. In this case, the authors have selected Geonames[2] because it is robust, comprehensive and frequently updated. A detailed explanation of the basic operation of this module can be found in [16].

There is an important problem related to this task: disambiguation. As an example, in the gazetteer can be found almost 20 different entries for cities and places called Madrid. Furthermore, if the locations are contained one within the other and both have the same designation, the system will always choose the most specific one. Also, news may contain many locations, but probably the events occur in only one place. KGNR can set the focus on a place that is not reflected in the news, when the scores for the different locations do not reach the acceptance threshold established. That means that the text is talking about several places so unimportant that the focus of the document is actually the place which includes all of them (a province, a region, or a country).

In summary, the system first extracts a set of named entities present in the news. For each found named entity, KGNR searches if it is included in a gazetteer; if not, it removes it from the possible locations. Then, it expands and filters the location to obtain all the places whose description matches the named entity in the news text using disambiguation mechanisms. Next, for each possible location, it calculates a weight and if it does not pass a minimum threshold value it is removed from the candidate list and transferred to a list of places discarded. In case it has not found any relevant place, then it tries to find a focus from this list. If that list is empty, then the location may be implicit, and it tries to solve the problem using OWL ontologies that store rules of inference related to city locations, or databases that store city maps of locations, tourist guides, etc. Otherwise, it finds a list of relevant places and sorts it according to their weight. Finally, it assigns to each location its appropriate thesaurus. This geographical information is used then to make recommendations to users depending on their present location. These recommendations vary with the change of location of the user.

Knowledge Generation module: With the information of the categorization obtained in the previous modules, and by conducting a comprehensive lexico-syntactic text analysis of the news, this module of KGNR gets a set of assertions in the form *subject-verb-predicate* [17] whose union and simplification establish a representation of the news in a topic map format. The methodology applied on this module is close to the explained in [18]. This representation is stored in an XTM standard format as an attribute over the news. This topic map will be used later in the recommendation system.

Recommendation module: This module is inspired in previous research works like [19], [20], [21]. As users provide information to the system by selecting their favorite categories, by making searches, and by visualizing or evaluating them, KGNR stores these data and sends them to the server. Each time a user accesses news, the recommendation module obtains the XTM associated file of the news and incorporates it to the user profile, combining it with the previous XTM of the news viewed by the user. So KGNR gets progressively an XTM of users with the knowledge of the topics that they find interesting (X+). As a user can also flag a piece of news as not interesting, this information is used to create another XTM of each user with information related to issues that are not attractive (X-). It can be seen graphically in Figure 2.

Thus, once the system has categorized a piece of news and it has generated its XTM, it compares it with the XTM of the users to obtain a numerical degree of similarity by using lexical and semantic similarity algorithms [22]. Taking into account other factors such as categories selected by the user (C), keywords used to make searches (K), the impact or relevance (R) of the news (related to its number of downloads, i.e. the number of times it has been read by users), and the distance from the current geographic location (G) with the geographical categorization of the piece of news (whose mechanism has

[2] http://www.geonames.org/

been previously explained), the function(1) to assign to the piece of news a recommendation score is obtained.

$$Rec = f(w_0S(X_n, X+, X-), w_1C_u, w_2K, w_3G, w_4R) \quad (1)$$

Where *Xn* is the XTM obtained of a piece of news; *S* is the method that calculates the similarity between topic maps; and *w0*, *w1*, *w2*, *w3* and *w4* are configurable system parameters [0,1] to weigh each of the factors. Finally, the result, *Rec*, is normalized between 0 and 1 and compared to a configurable threshold to decide whether or not the piece of news is sent automatically to the user's mobile terminal.

III. Case Study

In our research group web page[3] a sample video showing the basic functionality of the prototype can be found. The KGNR application allows users to configure their favourite topics, to perform searches using a basic search form, or by limiting the search to a specific topic, and to check the suitability of the recommendations received. Thanks to the use of a simulator, it is possible to emulate changes in the location of the user without the need to perform actual movements, so that it can be seen how KGNR changes its recommendations depending on that location. Attendees will be able to download the application, temporarily and free of charge, on their mobile terminal to try it. To test the system, a prototype has been developed in collaboration with the Heraldo Group. The authors have planned the test of the application with a set of 30 users and over 10,000 news per day, using 100 different categories. Currently, this test stage is under development.

IV. Conclusions

This paper presents a new approach that takes into account the geolocation of the users and their personal preferences to generate personalized news recommendations by combining NLP techniques, ontologies and classic recommendation methods. The system benefits of using topic maps, a simple but powerful form of knowledge representation. As future work, we plan, on the one hand, testing the process in more real environments, and on the other hand, verifying its integration with expert systems oriented to document management, as described in [23].

Acknowledgments

This research work has been supported by the CICYT project TIN2013-46238-C4-4-R and DGA-FSE. The authors also thank the Heraldo de Aragón.

References

[1] J. Liu, P. Dolan, and E. R. Pedersen, "Personalized news recommendation based on click behavior," in *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 2010, pp. 31–40.

[2] L. Li, D.-D. Wang, S.-Z. Zhu, and T. Li, "Personalized news recommendation: a review and an experimental investigation," *Journal of Computer Science and Technology*, vol. 26, no. 5, pp. 754–766, 2011.

[3] P. Lops, M. De Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender systems handbook*. Springer, 2011, pp. 73–105.

[4] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, p. 4, 2009.

[5] J. Bao, M. F. Mokbel, and C.-Y. Chow, "GeoFeed: A location aware news feed system," pp. 54–65, 2012.

[6] S. Pepper and G. Moore, "XML Topic Maps (XTM) 1.0-topicmaps. org specification," *TopicMaps. Org Authoring Group, Accessed*, vol. 23, no. 08, p. 2003, 2001.

[7] K. M. Markham, J. J. Mintzes, and M. G. Jones, "The concept map as a research and evaluation tool: Further evidence of validity," *Journal of research in science teaching*, vol. 31, no. 1, pp. 91–101, 1994.

[8] A. Garrido, O. Gómez, S. Ilarri, and E. Mena, "NASS: News Annotation Semantic System," in *23rd International Conference on Tools with Artificial Intelligence (ICTAI 2011)*. IEEE, 2011, pp. 904–905.

[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," in *Information Processing and Management*, vol. 24, no. 5. Pergamon Press, Inc., 1988, pp. 513–523.

[10] G. Siolas and F. d'Alché Buc, "Support vector machines based on a semantic kernel for text categorization," in *IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000)*, vol. 5. IEEE, 2000, pp. 205–209.

[11] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[12] A. L. Garrido, O. Gómez, S. Ilarri, and E. Mena, "An experience developing a semantic annotation system in a media group," in *Natural Language Processing and Information Systems*. Springer, 2012, pp. 333–338.

[13] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *27th International Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM, 2004, pp. 273–280.

[14] S. Sekine and E. Ranchhod, *Named Entities: Recognition, Classification and Use*. John Benjamins, 2009.

[15] L. L. Hill, "Core elements of digital gazetteers: placenames, categories, and footprints," in *Research and Advanced Technology for Digital Libraries*. Springer, 2000, pp. 280–290.

[16] A. Garrido, M. G. Buey, S. Ilarri, and E. Mena, "GEO-NASS: A semantic tagging experience from geographical data on the media," in *Advances in Databases and Information Systems*. Springer, 2013, pp. 56–69.

[17] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, "Using predicate-argument structures for information extraction," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 8–15.

[18] A. L. Garrido, M. G. Buey, S. Escudero, S. Ilarri, E. Mena, and S. B. Silveira, "TM-Gen: A topic map generator from text documents," in *25th IEEE International Conference onTools with Artificial Intelligence (ICTAI), 2013*. IEEE, 2013, pp. 735–740.

[19] Z. Yu, W. W. Song, X. Zheng, and D. Chen, "A recommender system model combining trust with topic maps," in *Web Technologies and Applications*. Springer, 2013, pp. 208–219.

[20] A. L. Garrido, M. Soledad Pera, and S. Ilarri, "SOLE-R: A semantic and linguistic approach for book recommendations," in *14th IEEE International Conference on Advanced Learning Technologies (ICALT), 2014*. IEEE, 2014, pp. 524–528.

[21] W. Chen and R. Persen, "A recommender system for collaborative knowledge." in *Artificial Intelligence in Education*, 2009, pp. 309–316.

[22] L. Maicher and H. F. Witschel, "Merging of distributed topic maps based on the subject identity measure (SIM) approach," *Proceedings of Berliner XML tags*, vol. 4, pp. 301–307, 2004.

[23] A. L. Garrido, A. Peiro, and S. Ilarri, "Hypatia: An expert system proposal for documentation departments," in *IEEE 12th International Symposium on Intelligent Systems and Informatics (SISY), 2014*. IEEE, 2014, pp. 315–320.

[3] http://sid.cps.unizar.es/SEMANTICWEB/GENIE/Genie_Downloads.html