

Ontology-based Multilingual Search in Recommendation Systems

Xuan Hau Pham¹, Jason J. Jung^{2*}, Ngoc Thanh Nguyen³, Pankoo Kim⁴

¹QuangBinh University, Vietnam

²Chung-Ang University, Korea

³Wroclaw University of Technology, Poland, ngoc-thanh.nguyen@pwr.edu.pl

⁴Chosun University, Korea, pkkim@chosun.ac.kr

Abstract: The information on the web is not only published by an original language, but also expressed in many different languages. Almost recommendation systems also lack mechanisms to support users overcoming the language problem. In these systems, it is difficult to search a specific value (e.g., movie artist, movie title in movie domain) by using native language. In this paper, we present our approach to deal with this problem. We develop an ontology-based multilingual recommendation system using integrated data from Linked Open Data to support user with in different languages on movie domain. Multilingual Movie Recommendation System (MMRS) for searching as a case of study is developed. In this system, we illustrate a more comfortable and flexible implementation.

Keywords: multilingual entities; Linked Open Data; interlink; movie; recommendation system

1 Introduction

Nowadays, user acquire information, including attributes within various media (e.g., television, radio, news paper, blog, and social networks) by a native language. The traditional recommendation systems cannot usually be applied, for efficiently searching the various media. Traditional systems have some (a few) languages to switch amongst, however, the languages have been obtained from translation machines. This leads to connection data between certain languages and other languages that is not easy. With the developed open data system, it allows connection multiple data in different languages.

In recommendation systems, most users face language problems. They want to search some content in either their native language or some other learned

* Corresponding author

languages. The information is often published on the web by original language and expressed in different languages. After publishing, it will be translated to different languages by themselves or a community. In fact, it is always difficult to search a specific entity when users do not remember the original name (e.g., English name), they only know your country name. For example, *Cameron, J.* is a famous director, in Korea some people want to search about him, but they have a problem, they do not remember his English name. How to search in this case? In this paper, we present our approach to deal with this problem.

Multilingualism becomes an important task in natural language processing. Multilingual systems have been designed either by the system (i.e., user has to switch among languages and number of languages is limited) or the community, group users (i.e., there are a lot of languages that are published). For example, Wikipedia¹ is a huge open data source that is edited and developed by a community of volunteers in the world. Its contents are described with in many different languages, including movie.

Multilingualism is an interesting topic on the web [5, 17]. Data will be integrated from multi-resources, associated with multilingual content. Each content is expressed in different languages. However, in this paper, we only take into account multilingual searches, based on integrated data based on LOD [2, 3, 8]. Linked Open Data (LOD), provides an effective mechanism to connect data from multiple data resources by using the Resource Description Framework (RDF) and the Hypertext Transfer Protocol (HTTP). We will extract data on a movie domain. There are several LOD of movies on the web and many other movie-related data (e.g., IMDB², LinkedMDB³, DBpedia⁴,...).

Our proposal focuses on the integrated information from multi-resources to put them into MMRS. It will help the user overcome the language problem. In this paper, we propose and develop a onotology-based multilingual search, to support the user when searching the entities in different languages in the system. Our approach is illustrated using “movie domain”. In this system, users can search a certain entity, within 145 different languages.

The outline of paper is organized as follows. In Sect. 2, we present related work about multilingual entities and ontology-based multilingual systems. In Sect. 3, we present integrated data based on LOD and ontology-based multilingual entity in recommendation systems. In Sect. 4, we develop our system and how it works. Finally, we talk about the major conclusions and future of our work herin.

¹ <http://www.wikipedia.org/>

² <http://www.imdb.com/>

³ <http://data.linkedmdb.org/>

⁴ <http://dbpedia.org/>

2 Related Work

Multilingual search is a challenge for social network (Facebook, Twitter, Flickr) and also open web (Wikipedia) [12, 7, 6] and recommendation systems [13, 19, 21]. Almost systems only take into account switching different languages by translating [4, 20]. The system quality depends on the translation quality. In the fact, as we know the translation machine makes a lot of mistakes. Thus, using multilingual entities based on ontology and integrated data from Linked Open Data (LOD) will improve the representation on the systems.

Ontology-based multilingual models are also proposed by [9, 15, 18]. In [9], authors proposed a new multilingual retrieval model, named CL-ESA. This model exploits the multilingual alignment of Wikipedia to represent a document as a concept vector and the similarity between two vectors can computed with the cosine similarity. In [15], the system, LabelTranlator, was proposed as an ontology to identify different languages labels. In [18], authors presented the translation-tree technique, that is based on an ontological representation for multilingual information retrieval. Each language is built as a multilingual ontology to map corresponding terms.

A multilingual search model for Flickr was proposed by Peinado et al. [7] and they called it *FlickLing*. This system allows to search monolingual and multilingual images and return a set of images with annotation in different languages. However, it can support six languages and applies a term-by-term translation for the multilingual search.

A fuzzy-based method for multilingual patent search, Fuzzy Logic Decision Support, is proposed by Segev et al. [11]. The patents are represented by a set of concepts related to a multilingual knowledge ontology. The model analyzed a several patents from Korean, US and Chinese as support for a multilingualism process.

The most important task in the multilingualism problem, is to extract name entity matching. In Wikipedia, the multilingual entity is represented as a set of InterLanguages-Links (ILL). A multilingual named entity recognition from Wikipedia is proposed by [6]. In this paper, the authors classify each article into “name entity”, in nine languages and project the links onto name entities. In [10], they introduce a fuzzy-based method to extract metadata automatically and cognitive metadata generation. They also apply different document parsing algorithms to extract rich metadata from multilingual content. This framework is evaluated on three languages, English, German and French. In order to measure the similarity between multilingual sentences, Adafre et al. investigated multilingual analysis to generate similar sentences in different languages [1].

The multilingual search does not only find related results that a user needs, but also returns new information for the user. It is also a challenge for novelty mining [14]. In our approach for multilingual search within recommendation systems; the

results can contain multilingual entities, where users can understand their content. For example, a user enters a query to find the director Cameron J., in Korean, the system will return a lot of information about this name in different languages (e.i., not only Korean but also other languages such as Japanese, German, Vietnamese and so on) and user can get several information.

In [8], we have applied integrated data from LOD to recommend not only a movie domain, but also books and music. The representation data on recommendation systems in different languages that are based on integrated data from LOD, will be better and more flexible, than traditional systems.

3 Ontology-based Multilingual Recommendation Systems for Searching

3.1 Multilingual Concepts and Integrated Data on Movie Domain

The main contribution of this paper is to propose a multilingual search process to match a search entity to its corresponding concept in different languages. In our approach, we try to implement the system in a movie domain. User do not need to remember the original names of any artist or any title (e.g., English name). User can enter their language and search a certain value.

Bilingualism and multilingualism are being discussed and developed within social networks and economic systems. For movie domains, they also try to fully support users. Most systems are monolingual, such as IMDB, LinkedMDB, MusicBrainz, etc. and a few are multilingual systems, such as, DBpedia, Wikipedia.

IMDB is a huge movie data repository. It describes a vast number of movies, with full information (e.g., title, genre, actor, director, music, URI, company, rating, runtime, and so on). Each entity is accessed by using URI, for example, the link <http://www.imdb.com/name/nm0000116> describes “James Cameron” director and <http://www.imdb.com/title/tt0499549/> is an URI that describes about “Avatar” movie.

LinkedMDB describes movie information based on movie entities, namely interlink. Movie entities have been extracted from IMDB, DBpedia and other sources.

DBpedia is open dataset on the Internet. It is organized by categories (e.g., movie, book, music and so on). Its information is extracted from Wikipedia. Users can easily access data by using the interlinks. In DBpedia, the entities matching are based on its properties. Table 1 shows properties on movie domain. For example, in order to find the matching between two entities, on actor or director, we have to

take into account *dbpedia-owl:starring* or *dbpedia-owl:director*. The name of *director* or *starring* will be represented as a list of entities in different languages.

Table 1
The properties on DBpedia and LinkedMDB for movie domain

DBpedia		LinkedMDB	
Property	Attribute	Property	Attribute
dbpedia-owl:work/runtime	Runtime	movie:actor	Starring
dbpedia-owl:abstract	Abstract	movie:cinematographer	Cinematographer
dbpedia-owl:cinematography	Cinematographer	dc:date	Event
dbpedia-owl:director	Director	movie:director	Director
dbpedia-owl:distributor	Distributor	movie:editor	Editor
dbpedia-owl:editing	Editing	movie:genre	Genre
dbpedia-owl:musicComposer	Composer	movie:initial_release_date	Release date
dbpedia-owl:producer	Producer	movie:language	Movie language
dbpedia-owl:writer	Writer	foaf:page	IMDB link
dbpedia-owl:starring	Starring	movie:producer	Producer
rdfs:comment	User comment	movie:runtime	Runtime
rdfs:label	Title	dc:title	Title
owl:sameAs	Multilingism	rdf:type	Type
dbpprop:language	Original language	movie:writer	Writer
dbpprop:country	Country	movie:actor_name	Actor name
dbpedia-owl:abstract	Abstract	movie:director_name	Director name
dbpedia-owl:alias	Alias	movie:film_genre_name	Genre name
dbpedia-owl:birthDate	Birthday	movie:editor_name	Editor name
dbpedia-owl:birthName	Name		
dbpedia-owl:birthPlace	Birthplace		
dbpedia-owl:birthYear	Birthyear		
dbpedia-owl:education	Education		
rdfs:comment	Comment		
foaf:givenName	Given name		
dbpprop:occupation	Occupation		

In our scenario, each entity (label, concept) will be detected by URI on the resource systems. It will identify the languages, contents, description and interlinks. In multilingual recommendation systems, each value of an item, as a concept, takes the information from resource data and has a list of corresponding different languages for the concept. This list is automatically detected and obtained from multilingual systems.

The following algorithm is used for integrating data:

Input: a list of movies, I

Output: a set of multilingual movie concepts

Algorithm:

Foreachi $\in I$

$V =$ a set of movie concepts from IMDB

Foreachv $\in V$

Mapping into LinkedMDB

Extracting data from DBpedia

Return $E =$ a set of multilingual movie concepts

Figure 1 shows the relationships between Korean information and English information of “The Spy” movie based on DBpedia properties (e.g., *dbpedia-owl:director*, *owl:sameAs*, *dbpedia-ko* and so on). We can see that each value of movie will be expressed by its corresponding properties. Since DBpedia data is extracted from Wikipedia, some contents are inconsistent. Thus, we use movie data from IMDB as standard information for integrating data.

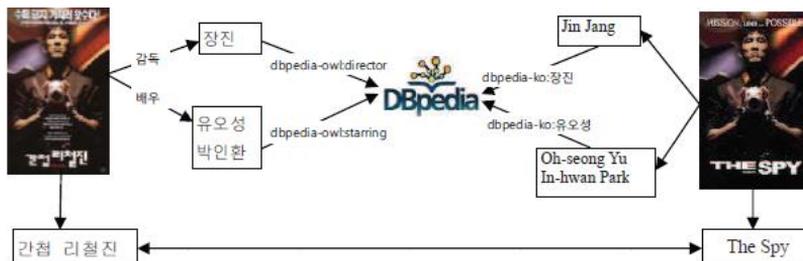


Figure 1
Multilingual entities matching

3.2 Multilingual Recommendation System

The aims of the recommendation system is not only suggest a set of new items based on user preference, but also explore the relationship among users and items. Therefore, multilingual recommendation systems will produce more interesting items in different languages and support users in overcoming the language problem. In order to understand our proposal we present some definitions for the multilingual concept and our model as follows:

Definition 1 (Multilingual Recommendation System Framework) A multilingual recommendation system is defined as a 6-tuple:

$$S = \langle U, I, A, V, L, R \rangle \quad (1)$$

where U is a set of users, I is a set of items, A is a set of attributes, V is a set of concepts, L is a set of languages and $R \subseteq A \times V \times L$ is a set of links. R can be represented as a set of interlink-languages (ILL).

Each $i \in I$, we can extract:

$$R_i = \{(a, v, l) | a \in A, v \in V, l \in L\} \quad (2)$$

For example, considering the movie *Titanic*, we obtain:

- (Title, “Titanic”, “en”, “[http://dbpedia.org/page/Titanic_\(1997_film\)](http://dbpedia.org/page/Titanic_(1997_film))”),
- (Direktor, “Cameron, J.”, “de”, “http://dbpedia.org/page/James_Cameron”),
- (Acteur, “LeonardoDiCaprio”, “fr”,
“http://dbpedia.org/page/Leonardo_DiCapri”),
- (Genre, “Adventure”, “en”, “http://data.linkedmdb.org/page/film_genre/31”).

LOD provides a mechanism to connect data from multiple resources. The connections can be established by links. In LOD, each concept is described by its content and interlink. We will use interlinks to find out the entities matching.

Definition 2 (Concept Matching) Let $v_1, v_2 \in V$ and $i \in I$, the matching between two concepts is computed as follows:

$$M(v_1, v_2) = \begin{cases} -1 & \text{if } v_1, v_2 \in R_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Searching on multilingual systems takes into account the entities matching. Each concept will have a set of different multilingual concepts which have the same description. For example, we consider the “Titanic” title, in other languages, in Table 2.

Table 2
The “Titanic” title in different languages

Title	Language	Interlink
Titanic	French	http://fr.dbpedia.org/resource/Titanic_(film,_1997)
Titanic	German	http://de.dbpedia.org/resource/Titanic_(1997)
타이타닉	Korean	http://ko.dbpedia.org/resource/타이타닉_(1997년_영화)
タイタニック	Japanese	http://ja.dbpedia.org/resource/タイタニック_(1997年の映画)
Titanic	Italian	http://it.dbpedia.org/resource/Titanic_(film_1997)

Each concept will have a set of corresponding entities. A set of concepts will have a set of corresponding interlinks. It will help improve the matching.

Definition 3 (Multilingual Search)

Given v is a search entity, the result of this search, is represented as follows:

$$M(v) = \{(v', l, r) | v' \in V, l \in L, r \in R: Lex(v) \subset V\} \quad (4)$$

where $Lex(v)$ is a function to return a set of result concepts.

In recommendation systems, the most important task is to build the user profile (user preferences). Each user will record all of their interactions and all the information in a session. The systems will discover these data to understand what user needs and predict recommendation for next time.

Definition 4 (User Preference) Give certain $u \in U$, the ontological user profile in the recommendation system is expressed as follows:

$$f(u) = \{(v, l) | \forall v \in V, \exists l \in L: (v, l) \subset R\} \quad (5)$$

Definition 5 (User Similarity) Given two users $u_1, u_2 \in U$. T the similarity between two users based on user preference is defined as follows:

$$Sim(u_1, u_2) = \frac{sim(v_{u_1}, v_{u_2})_{f(u_1) \cap f(u_2)}}{sim(v_{u_1}, v_{u_2})_{f(u_1) \cup f(u_2)}} \quad (6)$$

4 Multilingual Search Movie Recommendation System: a case of study

In this paper, we propose multilingual searches in a movie recommendation system, as a case study. Figure 2 shows the main the interface of our system. It is easy to input a name (e.g., movie, artist) and search. The result will be represented related-movie blocks.

- Number of results
- List of results
- <name, language>
- Related-contents
(e.g. a list of movies for an artist, list of artists for a movie)
- Multilingual entities, languages and interlinks

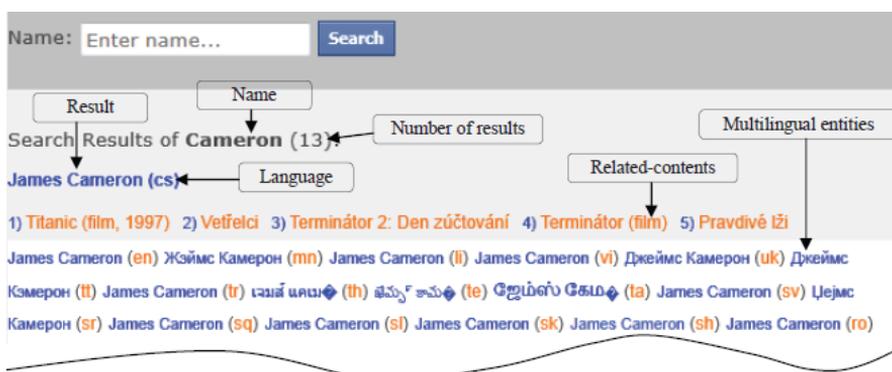


Figure 2

Main interface of multilingual movie search

In order to implement our system, we have extracted the dataset as Section 3.1. Table 3 expresses the multilingual entities in the system.

Table 3

The multilingual entities statistic in our system

Languages	#Artist	#Title
English	10845	1888
German	6614	1165
French	6877	1170
Russian	4546	998
Italian	5933	1100
Korean	3703	162
Japanese	4411	815
Vietnamese	795	61
Chinese	2237	281

Figure 3 shows the results of a search for the movie, “Titanic” in English, Korean and Russian languages.

We can extract the relationships among different languages for one movie. For example, we can extract 15 languages for the *Titanic* movie and 59 languages, for *Morgan Freeman*, the actor. When a user searches a certain name of movie, the system will show the movie information (e.g., director(s), actors, actresses) and a list of the same names in different languages (English (en), French (fr), German (de), Chinese (zh), Korean (ko), Japanese (ja), Vietnamese (vi), and so on). Also, when a user searches a certain name of an artist, it will show a list of movies that they were in and a list of the same their names. Figure 4 shows the connection between *Titanic* in English and *Cameron, J.* director, in Korean on multilingual movie search.

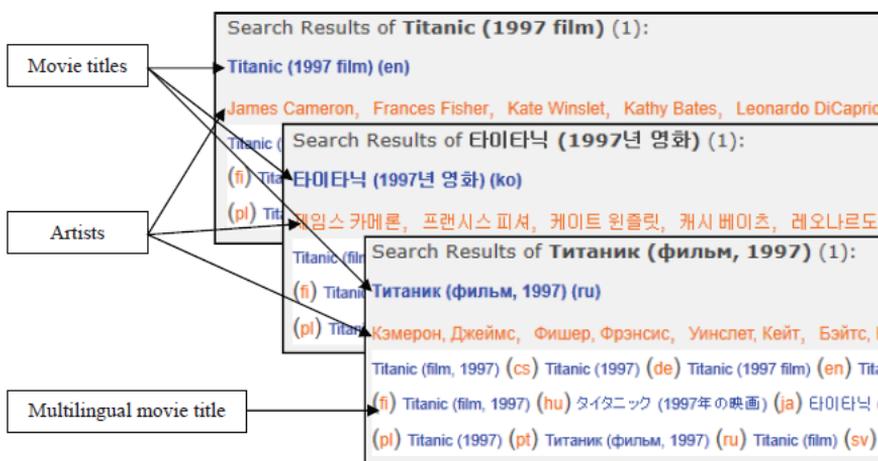


Figure 3
Representing “Titanic” movie on multilingual movie



Figure 4
The connections between multilingual entities on multilingual search

When a user searches a movie title, the system will return a set of movies including artists and multilingual titles. It will return a set of artist names in multiple languages and a set of movies that this artist was involved in when the user searches movie artist.

In order to develop multilingual search systems, that we have implemented on a movie domain. Movie information is extracted from IMDB and DBpedia. The quality and number of languages for multilingualism depend on data in those sources. It means that if the entities are well-known, then the multilingual entities will be more accurate.

In addition, the system will build user profiles for the various users. The recommendation processing is based on an ontological user preference, to predict

which is a better search item for each user. Figure 5 shows the recommendation interface in 3 languages (English, German, Vietnamese).



Figure 5

The recommendation interface

Concluding remarks

Multilingual searches in a movie recommendation system will bring a more flexible interaction for users. Users can overcome language problems. Each item not only is described on bilingual data, but also expressed in various languages. In this paper, we presented our approach to discover the relationships among multilingual concepts for searching on a movie domain and ontological user preferences are also considered in recommendation processing. We also developed a demo system for our proposal. However, the integrated data, based on LOD, is extracted offline and several languages are not available in the data resources. Thus, returned results are not a full expression.

As for future work, we will increase the number of entities and the number of movies. We would also like to show a comparison with other current approaches.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2013.12. Also, this work was supported by the Human Resource Training Program for Regional Innovation and Creativity through the Ministry of Education and National Research Foundation of Korea (NRF-2014H1C1A1066494). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2014R1A2A2A05007154).

References

- [1] Adafre S. F., and De Rijke M.: Finding Similar Sentences across Multiple Languages in Wikipedia, Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 62-69
- [2] Bizer C., Heath T. and Berners-Lee T.: Linked Data - the Story so Far, International Journal on Semantic Web and Information Systems, Vol. 5 (3), 2009, pp. 1-22
- [3] Hassanzadeh O. and Consens MP.: Linked Movie Data Base, Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW 2009), Spain, 2009
- [4] Hillier M.: The Role of Cultural Context in Multilingual Website Usability, Electronic Commerce Research and Applications, Vol. 2 (1), 2003, pp. 2-14
- [5] Jung J. J.: Cross-Lingual Query Expansion in Multilingual Folksonomies: A Case Study on Flickr, Knowledge-based Systems, Vol. 42, 2013, pp. 60-67
- [6] Nothman J., Ringland N., Radford W., Murphy T. and Curran J. R.: Learning Multilingual Named Entity Recognition from Wikipedia, Artificial Intelligence Vol. 194, 2013, pp. 151-175
- [7] Peinado V., Artiles J., Gonzalo J., Barker E. and López-Ostenero F.: Flickling: a Multilingual Search Interface for Flickr, Proceedings of CLEF 2008 Workshop Notes, Aarhus, Denmark, 2008
- [8] Bello-Orgaz G., Jung J. J., Camacho D.: Social Big Data: Recent Achievements and New Challenges, Information Fusion, Vol. 28, 2016, pp. 45-59
- [9] Potthast M., Stein B. and Anderka M.: A wikipedia-based Multilingual Retrieval Model, Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, ECIR'08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 522-530
- [10] Sah M., and Wade V.: Automatic Metadata Mining from Multilingual Enterprise Content, Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 11, 2012, pp. 41-62
- [11] Nguyen D. T., Jung J. E.: Real-Time Event Detection on Social Data Stream, Mobile Networks and Applications, Vol. 20 (4), 2015, pp. 475-486
- [12] Yarowsky D., Ngai G., and Wicentowski R.: Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora, Proceedings of the First International Conference on Human Language Technology Research, HLT '01, Association for Computational Linguistics, Stroudsburg, PA, USA, 2001, pp. 1-8

-
- [13] Lops P., Musto C., Narducci F., De Gemmis M., Basile P., and Semeraro G.: Mars: a Multilanguage Recommender System, Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, ACM, pp. 24-31
- [14] Zhang Y., Tsai F. S., and Kwee A. T.: Multilingual Sentence Categorization and Novelty Mining, Information Processing and Management, Vol. 47 (5), 2011, pp. 667-675
- [15] Espinoza M., Gómez-Pérez A., and Mena E.: Enriching an Ontology with Multilingual Information, Springer Berlin Heidelberg, 2008, pp. 333-347
- [16] Embley D. W., Liddle S. W., Lonsdale D. W., and Tijerino Y.: Multilingual Ontologies for Cross-Language Information Extraction and Semantic Search. Conceptual Modeling–ER 2011, Springer Berlin Heidelberg, 2011, pp. 147-160
- [17] Gracia J., Montiel-Ponsoda E., Cimiano P., Gómez-Pérez A., Buitelaar P., and McCrae J.: Challenges for the Multilingual Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web, 2012, pp. 63-71
- [18] Guyot J., Radhouani S., and Falquet G.: Ontology-based Multilingual Information Retrieval, CLEF Workshop, Working Notes Multilingual Track, 2005, pp. 21-25
- [19] Luberg A., Järv P., Schoefegger K. and Tammet T.: Context-Aware and Multilingual Information Extraction for a Tourist Recommender System, Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, ACM, 2011
- [20] Zahed F., Van Pelt W. and Song J.: A Conceptual Framework for International Web Design, Professional Communication, IEEE Transactions, Vol. 44 (2), 2001, pp. 83-103
- [21] Pham X. H., and Jung J. J.: Recommendation System Based on Multilingual Entity Matching on Linked Open Data, Journal of Intelligent and Fuzzy Systems, Vol. 27(2), 2014, pp. 589-599