Enhancing Content Based Filtering Using Web of Data

Hanane Zitouni, Souham Meshoul, Kamel Taouche

Enhancing Content Based Filtering Using Web of Data

Hanane Zitouni[1], Souham Meshoul[1] and Kamel Taouche[2]

[1]Department of Computer Science, Abdelhamid Mehri University, Constantine, Algeria
[2] Claude Bernard Lyon 1 University, Lyon, France

hanane.zitouni@univ-constantine2.dz, souham.meshoul@univ-constantine2.dz, kamel.taouche@liris.cnrs.fr

Abstract. Recommender systems are very useful to help access to relevant information on the web and to customize search. Content based filtering (CBF) is an alternative among others used to design recommender systems by exploiting items' contents. Basically, they recommend items based on a comparison between the content of items and user profile. Usually, the content of an item is represented as a set of descriptors or terms; typically the words that occur in text documents. The user profile is represented by the same terms and built up by analysing the content of items he used before. However current CBF recommender systems are mostly devoted to deal with textual resources and cannot be used in their current form to handle the variety of data published on the web especially unstructured data. Another challenge for the existing CBF methods is the issue of new user for whom the system cannot draw any inference due to the lack of information about the user. This paper describes an approach to CBF that aims to deal with these problems on which CBF systems perform poorly. The basic feature of the proposed approach is to incorporate linked data cloud into the information filtering process using a semantic space vector model, and FOAF vocabulary, which is used to define a new distance measure between users, based on their FOAF profiles. We report on some experiments and very promising results of the proposed approach.

Keywords: **Content based filtering, RDF, linked data, FOAF vocabulary, Web of Data, SPARQL.**

1 Introduction

Nowadays, the explosive growth of data published on the web in all different fields such as e-learning, social networks, e-commerce among many others is not slowing down soon according to recent studies [1], [2]. The expanding data universe makes it difficult to get benefit from the web content. Furthermore, predicting user responses to options for recommendation purpose becomes an enormous challenge for an extensive class of web applications. Recommending a resource is usually achieved through information filtering. There exist two major approaches to information filtering [1]: Collaborative filtering and Content-based filtering. A Collaborative Filtering (CF) system chooses items based on the correlation between people with similar preferences, while

a content-Based Filtering system (CBF) selects items based on the correlation between the content of the items and the user preferences.

Despite the demonstrated effectiveness of CBF technology in many cases, some drawbacks make it inappropriate in its current form for other cases. Indeed, CBF requires analyzing the content of a document which is computationally expensive and even impossible to perform on multimedia items which do not contain descriptive text [3]. Furthermore, CBF presents difficulties to handle the new user problem where no preference is available. At the beginning, a new user does not have any preference value. Therefore, it is very hard to issue any recommendation to him.

In this paper, we propose solving these issues by enhancing CBF systems using semantic derived from the *Web of Data*. In this latter, the World Wide Web is viewed as a global database by creating links between data which known as *Linked Data*. When these linked data enable describing people, they are called FOAF (*Friend of A Friend*).The proposed approach is based on Vector Space Modeling of CBF[4], and enhanced by a semantic level extracted from the web of data leading to a new model that we refer to as *Semantic Vector Space Model* (SVSM).

Following this introduction, CBF based on Vector Space Model is described in section 2.Section 3 presents key features of the web of data. In section 4, a review of some related works that propose recommender systems in web of data context is given. In section 5, we describe the proposed approach SCBF and we report on the conducted experimental study and obtained results. Finally, conclusion and future work are given.

2 Content Based Filtering (CBF)

Information filtering deals with the delivery of information that would be interesting and useful to a user given his profile and preferences. An information filtering system assists users by filtering the data source and deliver relevant information to them. When the delivered information comes in the form of suggestions such information filtering system is called a recommender system. A CBF technique, also referred to as cognitive filtering [1], recommends items based on a correlation between the content of the items and a user profile. The content of each item is represented as a set of descriptors or terms, classically the words that occur in a document. The user profile is represented by a set of terms built up by analyzing the content of items seen by the user. Typically, a content based filtering system selects relevant items based on the correlation between the content of the items and the user's preferences.

One of the most important approaches is *Vector Space Model (VSM)* or term vector model [5]. In the vector space model, a document *D (item)* is represented as an m-dimensional vector, where each dimension corresponds to a distinct term [6]. The term frequency (tf) is a numerical statistic that measures the importance a term would have with regard to a document in a collection or corpus:

$$\mathrm{tf_{vi}} = \frac{\mathrm{n_{vi}}}{\mathrm{N}} \quad (\mathbf{1})$$

Where, n_{vi} is the number of times term t_i appears in a vector *v*; it models the taste of user and *N* is the total number of terms in the vector *v*.

To measure the extent to which documents contain a given term t_i we need to calculate the inverse document frequency (*idf*).

$$\mathrm{idf_i} = \log\left(\frac{\mathrm{D}}{\mathrm{n_j}}\right) \quad (2)$$

Where, D is the total number of documents, n_j is the number of documents d_j containning term t_i.

From *tf* and *idf* we can calculate the *weight* (*W*) or *tfidf*. This latter is a concept that can be used to create a profile of an item for example a document or an object…etc.

$$\mathrm{W_i} = \mathrm{tf_{vi}} * \mathrm{idf_i} \quad (3)$$

A content-based filtering system selects relevant items based on the correlation between the content of the items and the user's preferences [7]. However this technique suffers too from some disadvantages such as: it requires analyzing the content of the document which is expensive and even impossible to perform on multimedia [8] and the problem of new user or no preferences problem. At the beginning, a new user does not have any preference values; this makes it impossible to give him any recommendation. To address these problems, we propose to enhance CBF using the *Web of Data*.

3 Web of Data

Typically, a data set published in the web contains knowledge about a particular domain, like books, music, encyclopedic data and companies to name just few. If these data sets were interconnected i.e. linked to each other, this makes the World Wide Web a global database termed by Tim Berners Lee as *Web of Data.*

The most important concepts related to the web of data are: *Linked Open Data (LOD), Friend of A Friend (FOAF) vocabulary,* and *Resource Description Frame work (RDF).*

3.1 Linked Open Data

The term Linked Open Data refers to a set of best practices for publishing and connecting structured data on the web using international standards of the World Wide Web Consortium.[1] LOD cloud is considered as a network or collection of data silos.

The diagram of figure1 is maintained by Richard Cyganiak, and Anja Jentzsch (http://lod-cloud.net/).

The core of this diagram is *DBpedia*[2] which is a community effort to extract structured information from Wikipedia and to make this information available on the Web.

[1] http://www.w3.org/standards/

[2] http://www.wiki.dbpedia.org/

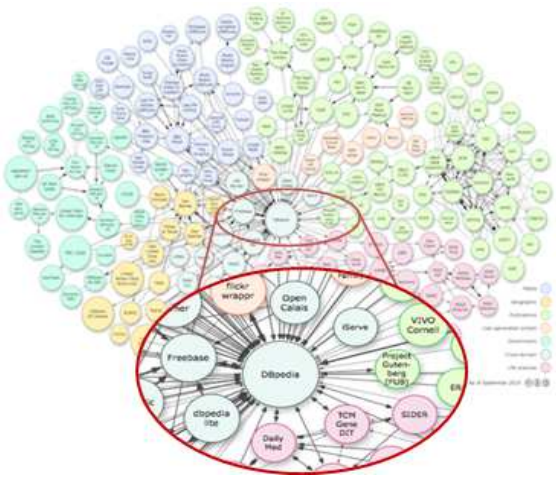

Fig. 1. Liked Data Cloud Diagram [9].

3.2 FOAF Vocabulary

The FOAF project began as an "experimental linked information project." Dan Brickley and Libby Miller are responsible for its inception, and EddDumbill and Leigh Dodds (http://www.foaf-project.org) notably contributed to its success. FOAF enables to describe people, their interests, their achievements, their activities, and their relationship with other people [9]. In Table 1 below all FOAF classes and proprieties are presented**.**

Table 1. Classes and proprieties of FOAF vocabulary.

Classes		
Agent	Document	Group
Image	LabelProprety	OnlineAccount
OnlineChatAccount	OnlineEcommerceAccount	OnlineGamingAccount
Organization	Person	PersonalProfileDocument
Project		
Proprieties		
Account	AccountName	AccountServiceHomepage
Age	AimChatID	Based near
Birthday	CurrentProject	Depiction
Depicts	DnaChechsum	FamilyName
First name	FirstName	Focus
Funded	Geekcode	Gender

GivenName	Givenname	HoldsAccount
Homepage	IcqChatID	Interest
IsPrimaryTopicOF	JabberID	Knows
LastName	Logo	Made
Maker	Mbox	Mbox sha1sum
Member	MembershipClass	MsnChatID
MyersBriggs	Name	Nick
Openid	Page	PastProject
Phone	Plan	PrimaryTopic
Publications	SchoolHopage	Sha1
SkypeID	Status	Surname
Theme	Thumbnail	Tipjar
Title	Tobic	Topic interst
Weblog	WorkInfoHomepage	WorkPlaceHomepage
YahooChatID		

3.3 Resource Description Framework (RDF)

Resource Description Framework or in short RDF provides a common data model for Linked Data [9] and is particularly suited for representing data on the Web. Linked Data uses RDF as its data model and represents it in one of several syntaxes. There is also a standard query language called *SPARQL*. A single RDF statement describes two things and a relationship between them. Technically, this is called an *Entity-Attribute-Value (EAV)* data model.

4 Related Work

Few recommender systems based on web of data have been developed till date. The following table 2 reviews some recent approaches. It provides a short description of the methods and indicates the web of data concepts used.

Table 2. Recent recommender systems enhanced by web of data.

Systems	Technique	Recommendation domain	Using FOAF Vocabulary	Using Linked Open Data	Description
Foafing the Music [10]	CBF	Music	Yes	No	It used RDF data. In order to provide its recommendations it crawled data from a large number of web sites.

Mining Recommendations[11]	CF	Unspecified	No	No	Uses the new metric to compare the recommendations that were generated from the public web
Music-related recommender Systems[12]	CBF	Music	No	Yes	A Content-based recommendation is discussed, which uses not only meta-data but also the audio signal of the music for recommendations.
Linked data recommendation[13]	CF	Music	No	Yes	It presents how to exploit the benefits of the LOD community effort to build recommender systems. By providing public, collaboratively created and semantically structured data
SPrank[14]	CF	Unspecified	No	Yes	Semantic Path-based Ranking SPrank a novel hybrid recommendation algorithm able to compute top-N item recommendations from implicit feedback exploiting the information available in the so called Web of Data.
LOD for content-based recommender systems[15]	CBF	Movies	No	Yes	It implemented a content-based RS that leverages on the data available within Linked Open Data DBpedia and LinkedMDB datasets in order to recommend movies to the end users.
Movie Recommender system [16]	Hybrid	Movies	No	No	Investigate the use of folksonomies to generate tag-clouds that can be used to build better user profiles to enhance the movie recommendation. They use an ontology to integrate both IMDB and Netflix data.
Linked Data Datasets for sameAs Interlinking Using Recommendation Techniques[17]	CF	Unspecified	No	Yes	In this work they to treat

From Table 2, we can observe that most proposed approaches are dedicated to a specific domain example movies or music and use either FOAF vocabulary or linked data cloud. Almost half of these methods are based on Collaborative filtering.

Our work is motivated by the fact that combination of FOAF vocabulary and linked data cloud would have the potential to further improve the ability of CBF to achieve suitable recommendations. Using the FOAF vocabulary helps in solving the problem of new user and the extracted linked data from the cloud provide a semantic description of non-structured items.

5 Proposed Semantic Content Based Filtering (SCBF)

CBF selects items based on the correlation between the content of the items and the user's preferences. As aforementioned, the problem with CBF is that it requires analyzing the content of the items which is expensive or impossible with multimedia items. To solve this issue along with the new user problem, we describe in this section how the *Web of Data* technologies could be used to enhance CBF systems. We refer to the proposed web of data based variant of CBF as Semantic Content Based Filtering (SCBF). In SCBF, we suggest integration of the following technologies:

- ***FOAF Vocabulary:*** if new user is connected, his FOAF description will be compared with the other users' FOAF descriptions. The comparison is based on the proposed formula :

$$\mathrm{D_{FOAF}(u,v)} = 1 + \log\left(\frac{1+\mathrm{K}}{\mathrm{P}}\right) \quad (4)$$

  Where, $D_{FOAF}(u, v)$ is the FOAF distance between users *u* and *v*, $K = L + S$ with *S* is number of the similar FOAF proprieties between users *u* and *v*, *L* is number of links between *u* and *v* and *P* stands for the total number of FOAF proprieties describing target user *u*.
  Following some important properties for the *class person [9]*:

  - *Based near* - A location that something is based near, for some broadly human notion of near (The *based near* relationship relates two "spatial things").
  - *Age* - The age in years of some person.
  - *Gender* - The gender of this person (typically but not necessarily 'male' or 'female').
  - *Title* - Title (Mr, Mrs, Ms, Dr. etc).
  - *Knows* - A person known by this person (indicating some level of reciprocated interaction between the parties).
  - *dMaker* - An agent that made this thing.
  - *Member* - Indicates a member of a Group.
  - *Interest* - A page about a topic of interest to this person.
  - *Topic_interest* - A thing of interest to this person.

- ***Linked Data Cloud***
  The vector space model is a representation often used for *text items* In this model, an item *i* is represented as an m-dimensional vector, where each

dimension corresponds to a distinct term. However, this technique is too limited with unstructured and even with semi-structured items.
To fix this problem, we propose in SFBC to enhance the *m* dimensional vector by *n* other textual or semantic attributes extracted from the linked data cloud. Therefore, the representation of the item will include (*m*+*n*) attributes and expressed of a (*m*+*n*)- dimensional vector that we refer to as *Semantic Vector Space Model (SVSM)*. The example below brings more explanation about the proposed SVSM.
In the dataset Movielens[3], the movie "*No escape*" is represented by the following textual attributes:

Id	**Title**	**Realise date**	**Genre**
1416	No escape	1994-01-01	Action, Science Fiction

On the same movie and using DBpedia, we can extract other information such as those given in the following Table 3.

Table 3. Textual and semantic attributes describing the movie "no escape".

Textual Attributes			*Semantic Attributes*		
Budget	**Director**	**Country**	**Language**	**Subject**	**image Size**
2.0E7	Martin Campbell	USA	English	Prison films	250

For that we propose a new version of the *tf* denoted by $\widetilde{tf}$ defined as follows:

$$\widetilde{\mathrm{tf}}\,(\mathrm{v},\mathrm{i}) = \frac{\mathrm{NS_{vi}}}{\mathrm{T}} \tag{5}$$

Where, $\mathrm{NS_{vi}}$ is the number of times triplet t_i appears in the semantic segment of the vector *v* and *T* is the total number of triplets in the semantic segment of the vector *v*.

$$\widetilde{\mathrm{idf_i}} = \log\left(\frac{\mathrm{Tt}}{\mathrm{n_j}}\right) \tag{6}$$

Where, Tt *is*the total number of triplet and n_j is the number of documents $d_j$where triplet $t_i \in d_j$.
Therefore, the semantic weight is given by:

$$\widetilde{W_t} = \widetilde{\mathrm{tf_{v,i}}} * \widetilde{\mathrm{idf_i}} \tag{7}$$

And the global weight *Wg* for the item is defined as follows:

$$\mathrm{Wg_i} = \mathrm{Wi} * \widetilde{W_t} \tag{8}$$

Based on the above description, the proposed SCBF approach suggests the following architecture of recommender systems shown on figure 2.

[3] https://movielens.org/

In the case of new user (the feedback is empty), his D_{FOAF} is calculated using other users, just after we recommend the set of items liked by the user who has the maximum D_{FOAF}.

The Space of attributes that describe the items is enhanced by semantic and textual attributes extracted from linked data cloud, which gives further descriptions of the items.

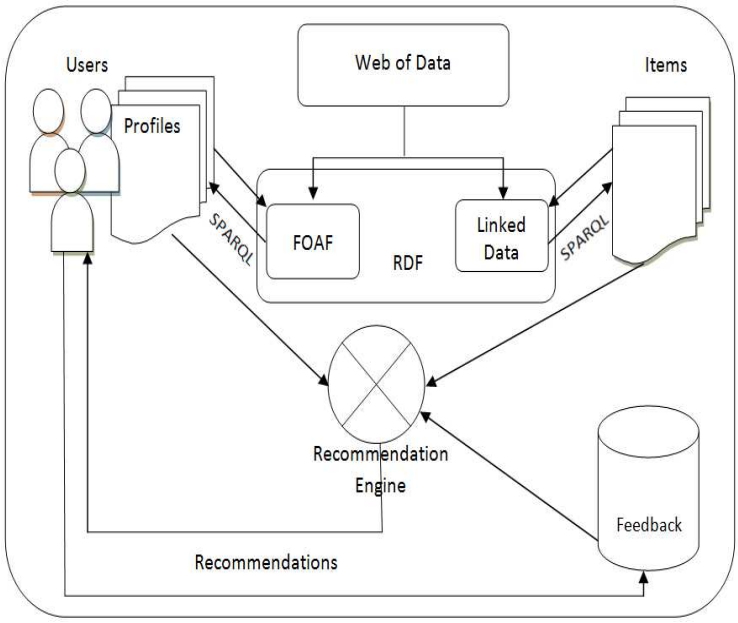

Fig. 2. General Architecture of SCBF.

The proposed SCBF engine can be outlined by the following algorithm.

Algorithm: Semantic Content Based Filtering Recommendation (SCBFR)

Input: U: set of users; I:set of items; F:Feedback; Triplet-RDF;
Output: List of recommendation;
1: **for** (i=1,…,Size(U)) **do**
2: **if** (F of U_i = = ∅) **then** /* in the case of a new user */
3: **for**(j=1 to Size(U)) **do**
4: Calculate $D_{FOAF}(U_i,U_j)$; /*based on the formula 4*/
5: **if** (Max< D_{FOAF}) **then**
6: Max= D_{FOAF};
7: Close_User= U_i;
8: **end**
9: **end**
10: Recommend list of items liked by Close_User to U_i;
11: Collect Ratings of U_i;
12: Create F of U_i;

```
13:      else
14:        for (n=1,…, Size(NI)) do /*NI are the None Rated Items by Ui*/
15:          Create vector v based on GRI  /*GNI are the Good Rated Items by Ui*/
16:          Calculate Wg for the item In;
17:          Select k items with the maximum value of Wg; /*based on the formulas
18:          6,7,8*/
19:          Collect Ratings of Ui;
20:          Update F of Ui;
21:        end
22:      end
     end
```

6. Experiments

In the dataset Movielens[4], all movies are characterized by the following attributes: *Id*, *Title*, *Realize date*, and *Genre*, using following SPARQL query based on the federation (released by FedX) [18], between DBpedia and Linked Movie DataBase (LDMDB[5]). We can extract more information about these movies like: *Director*, *Country*, *Actor,* and *Abstract*. The common attribute between Movielens and the federated query is the movies titles. Following is the SPARQL query that extract more information about Movielens movies:

```
SELECT * WHERE {?film <http://data.linkedmdb.org/resource/movie/filmid> ?uri.

      ?film <http://purl.org/dc/terms/title> ?Title .

      ?film <http://data.linkedmdb.org/resource/movie/actor> ?cast .

   ?cast <http://data.linkedmdb.org/resource/movie/actor_name> ?Actor .  ?film <http://data.linkedmdb.org/resource/movie/country> ?CountryID .

   ?CountryID <http://data.linkedmdb.org/resource/movie/country_name> ?Country .?film <http://purl.org/dc/terms/date> ?Year .

    ?film <http://data.linkedmdb.org/resource/movie/director> ?directorID .

    ?directorID <http://data.linkedmdb.org/resource/movie/director_name> ?Director.
         ?film <http://www.w3.org/2002/07/owl#sameAs> ?x.
      ?x <http://dbpedia.org/ontology/director>  ?director.?x <http://dbpedia.org/ontology/abstract>  ?abstact.}
```

[4] https://movielens.org/
[5] http://www.linkedmdb.org/

To measure the effectiveness of our approach, we calculated the Mean Absolute Error MAE, and Root Mean Square Error (RMSE) using the following formulas:

$$MAE=\frac{\sum_{u,i}|p_{u,i}-n_{u,i}|}{n} \quad (9)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{u,i}p_{u,i}-n_{u,i}^{2}} \quad (10)$$

Where $n_{u,i}$ is the note given by the user u on item I, $p_{u,i}$ is the predicted note, n is the total number of predicted notes.

The value MAE and RMSE of SBCF are compared with other values of state of the art techniques described in [19]. The results are shown on Table 4 where we can observe that the proposed approach offers the minimum error value.

Table 4. Comparative results.

	CBF	CF_U	CF_I	SCBF
MAE	0.83	0.86	0.83	0.46
RMSE	1.04	1.04	1.09	0.49

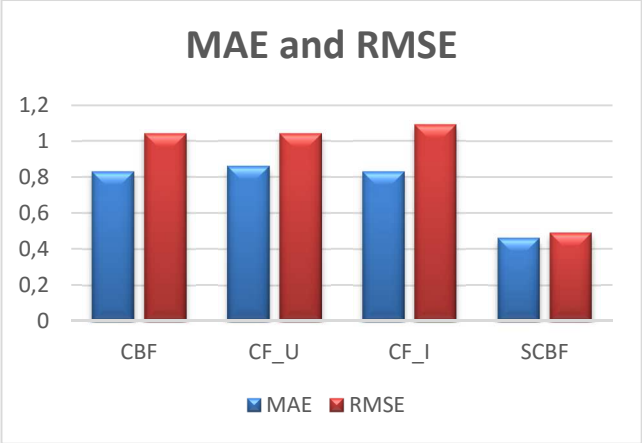

Fig. 3. Experimental Results of MAE and RMSE.

7. Conclusion

In this work, we described a new approach to content based recommendation using web of data which is mainly supported by some of intelligent technologies namely: FOAF vocabulary and Linked Data Cloud. We were faced with a challenge to use the technique of CBF while reducing the impact of new user issue and the difficulty of analyzing unstructured items. Promising preliminary results have been obtained. As future work, our plan is to test and evaluate the proposed approach with other

metrics like recall and precision, and apply new user problem solution to Collaborative Filtering (CF) algorithm to reduce the impact related to cold start issues.

References

1. BURKE, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction : (2002),331–370.
2. Zitouni, H., Nouali, O., and Meshoul, S. : Toward a New Recommender System Based on Multi-criteria Hybrid Information Filtering. In Computer Science and Its Applications (pp. 328-339). Springer International Publishing.R (2015).
3. Shoval, P., Maidel, V., Shapira, B.:An ontology- content-based filtering method Internationanl journal,vol.15, (2008).
4. Raghavan, V. V. and Wong, S. K. M. A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science, Vol.37 (5), (1986), p. 279-87,.
5. Salton, G., Wong, A., and Yang, C. S.: A Vector Space Model for Automatic Indexing, Communications of the ACM, vol. 18, nr. 11, (1975), pages 613–620.
6. Karen, S.J: A statistical interpretation of term specificity and its application in retrieval , Journal of Documentation, vol. 28, no 1, (1972), p. 11–21.
7. Shoval, P., Maidel, V., Shapira, B. :An ontology- content-based filtering method Internationanl journal,vol.15, (2008).
8. Dai, H., Mobasher, B., :Using ontologies to discover domain-level web usage profiles, Proc. of the Second Semantic Web Mining Workshop at PKDD 2001, Helsinki, Finland.
9. Wood, D., Zaidman, M., Ruth, L., Hausenblas, M.: Linked Data, by Manning Publications, (2014).
10. Celma, O., and Serra, X.. Foafing the music: Bridging the semantic gap in music recommendation. Web Semantics: Science, Services and Agents on the World WideWeb, (2008) 250–256.
11. Shani, G.; Chickering, M.; and Meek, C.. Mining recommendations from the web. In ACM Conference on Recommender Systems,. ACM New York, NY, USA,(2008), 35–42 .
12. Passant, A., and Raimond, Y.. Combining Social Music and Semantic Web for music-related recommender systems. In Social Data on the Web Workshop, (2008).
13. Passant, A., Heitmann, B., & Hayes, C.. Using linked data to build recommender systems. Proceedings of RecSys, New York, USA, (2009).
14. Ostuni, V. C., Di Noia, T., Di Sciascio, E., & Mirizzi, R.. Top-n recommendations from implicit feedback leveraging linked open data. InProceedings of the 7th ACM conference on Recommender systems (2013). ACM, pp. 85-92.
15. Mirizzi, R., Di Noia, T., Ostuni, V. C., & Ragone, A.. Linked Open Data for content-based recommender systems, (2012).
16. Szomszor, M., Cattuto, C., Alani, H., O'Hara, K., Baldassarri, A., Loreto, V., & Servedio, V. D.. Folksonomies, the semantic web, and movie recommendation, (2007)
17. Liu, .H , Wang, .T, and Tang, .T, al : Identifying Linked Data Datasets for sameAs Interlinking Using Recommendation Techniques, springer, (2016), pp 298-309
18. Schwarte, A., Haase, P., Hose, K., Schenkel, R., & Schmidt, M. Fedx: Optimization techniques for federated query processing on linked data. In International Semantic Web Conference, (2011), pp. 601-616. Springer Berlin Heidelberg.
19. Haddad, M. R., Baazaoui, H., Ziou, D., & Ghézala, H. B. Un modèle de recommandation contextuel pour la prédiction des intérêts des consommateurs sur le Web. IC2015, (2015).