Relevance Ranking for Services Retrieval

[1]Ouail Ouchetto, [2]Hassania Ouchetto and [2]Ounsa Roudies
[1]Department Mathemtics and Computer Science,
Faculty of SJES, Hassan II Ain Chok University, Casablanca, Morocco
[2]Department Computer Science,
Mohammadia School of Engineers, Mohammed Vth University-Agdal, Rabat, Morocco

Absract: Problem statement: One of the challenges of e-gov systems is to provide, during a search process, relevant services that meet user expectations. Indeed, obtaining relevant information responding to user queries is a difficult process. It becomes even complex when the query terms have many meanings and do not fit with the vocabulary used by the services. **Approach:** We propose an appropriate method to assess the adequacy of rendered services. This new method is based on a mathematical representation. It is based on calculating the relevance weight of each service by using the semantic equivalence. **Results:** Validation of this method was done in two times. Initially, it was implemented and integrated in a retrieval system. In a second step, it was made available to a number of users to give their judgment. **Conclusion:** The experiments show a high level of satisfaction of this method by improving the quality of the relevance ranking. The relevant services are presented in the first page and the order of relevance decreases with the pages.

Key words: Relevance ranking, services retrieval, semantic layer

INTRODUCTION

For the computer industry, the personalization of information is a major issue in the context of enterprise information systems, electronic commerce, electronic government and the knowledge access. The relevance of the provided information, its intelligibility and its adaptation to use and user preferences constitute the factors that determine the success of implementing such systems.

In the e-government systems, access to information and to the relevant services which fit to both the user context and user requirements represents a huge challenge for governments. This is due to many factors: the complexity of these systems, the diversity of the actors involved in the search process and the proliferation of heterogeneous resources constituting these systems (structured data, text documents, components). Therefore, information's diversity and user's disorientation are the main reasons of non-user's satisfaction of e-government services during a search process (Ouchetto *et al.*, 2012).

In a process of searching an e-gov service or information, the user information needs are often expressed by using some keywords and short phrases. Different query terms can be used to retrieve services. However, the user often does not build a query which accurately reflects his needs because of: (i) the user perspectives and terminological habits (ii) the difficulty of formulating a query, (iii) non-mastery of the vocabulary used by e-gov services and (iv) control's lack of the user's real needs who prefers to look in the long results' lists which do not meet his/her expectations than to look for the appropriate keywords.

To resolve this problem, we believe that the integration of a method for evaluating the services' appropriateness, as an important element in the search process of e-gov services, becomes an absolute necessity. The assistance to be brought is related to the final presentation of services. The principle is as follow: the user starts the search with a fixed need and a specific context, the system takes the keywords of the query. It enriches the query by including the semantics of these keywords. Afterwards, it calculates the weight of the retrieved services. Finally, it orders them in descending order before presenting them to the user.

Related work: Research communities in the field of information retrieval believe that relevance is a strategic point in all personalization systems. Its purpose is to make information relevant to the user. To achieve this goal, they developed several methods to improve the user's query, based on additional knowledge of the user. These methods are complemented by query expansion algorithms to remove the ambiguity of the meaning of terms used in the user's query (Bhogal *et al.*, 2007).

Corresponding Author: Ouail Ouchetto, Department Mathemtics and Computer Science, Faculty of SJES, Hassan II Ain Chok University, Casablanca, Morocco

All definitions of relevance have a common point which is the dependence between the information given and the users' needs expressed as a query. In most cases this query does not reflect the real requirements of the user. For Robertson *et al*. (1982), a document is relevant if it matches the user's needs in terms of information retrieval. It is called irrelevant if the user does not want it. However, Rijsbergen (1979) mentioned that relevance is a very subjective notion. Indeed, what is considered by someone relevant may not be by others.

According to Wallis and Thom (1996), a document is relevant if it satisfies certain requirements which are implicitly defined in the user mind. They emphasize the importance of controlling the relevance definition. In fact, the user must differentiate between what is relevant and what is not. In addition, the user needs are usually different from what he describes. Actually, users do not express clearly their needs and consequently, they can express the same need by using different queries with very different meanings.

To provide users with relevant information corresponding to their needs and expectations, the process of information retrieval must be based on a model of relevance. When a user enters a query, this model allows the calculation of the relevance of each retrieved information. Those who have the best relevance score will then be presented to users in descending order. We'll talk about calculating "ranking".

Methods of automatic indexing for texts were developed in the 1960s. They implemented the approach of bag-of-words which still exists until this time. Even though automatic indexing is widely used today, many information providers and even services available online, still count on the human effort to obtain the relevance information.

In the 1970s, research has been oriented to partial match retrieval models. Thus, the probabilistic models were developed. However, it was not until the 1990s that partial match models were able to succeed in the market through the Web development and search engines. This model applies probability theory to information retrieval systems. It is based on two principles (Kowalski and Maybury, 2000): (i) "The most promising source of techniques for estimating the usefulness of probabilities for output ranking in IR is standard probability theory and statistics". (ii) "If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data".

Another type of document relevance calculating method is based on lexical cohesion with structure analysis. In this method, documents are formalized with lexicon chains that are constructed by extracting semantic clusters of words by using the semantic dictionary HowNet, then the weight of each lexical chain is evaluated and finally. The relevance of documents is calculated with their performances (Yu-Ming *et al*., 2008).

There are other methods based on the user's profile. The later contains relevant information about users, such as interests and personal preferences. They play a role and are key to personalization. Mianowska and Nguyen (2011) proposed a method of simulating the behavior of users and takes into account the user's profile to improve the relevance of the results. Indeed, they proposed an algorithm for judgment of relevance based on user preferences.

However, the acquisition of user profiles in an efficient way remains a challenge. Several techniques have been proposed for the collection of information (Middleton *et al*., 2004; Ouchetto *et al*., 2011). These techniques can be classified in three types: questionnaire, feedback and, user's interactions. The techniques which are based on the questionnaire ask users to complete some given forms. In the techniques based on the users feedback, the users have to make their judgment about information relevance according to their needs (Robertson and Soboroff, 2002). However, these mechanisms have shown their ineffectiveness. They are very uncomfortable for the user As (Sugiyama *et al*., 2004). The third type of techniques does not involve the user. Information is collected in a transparent way from all the historical interactions and navigations (Gauch *et al*., 2003; Liu *et al*., 2004). In this case, user profiles may contain inaccurate information. The user's behavior can be unpredictable and his search can be varied and random. Indeed, they can occur in areas of every type and kind which are neither part of its interests or preferences.

In this context, the question which seems to be reasonable is: how to calculate information relevance without taking into account the user profile? We will try to answer this question in this study by proposing a new method. Subsequently, we will evaluate the proposed method on a descriptive basis of dedicated service to the field of e-gov. This method has several advantages. It depends on the user's profile and therefore we do not care about complex mechanisms for managing user profiles and their update. Its integration in the retrieval system is spontaneous. It is to be noted that this approach is applied in the context of a search system incorporating a semantic layer.

MATERIALS AND METHODS

Services descriptive base and ontology: In this study, we propose a method for calculating and evaluating the relevance between the users information needs and the retrieved e-gov services. The services are stored in a Services' Descriptive Base (SDB). In the basis, a service is described by a set of dimensions that can be elemental or composite (Fig. 1). The dimensions of e-gov services are: Beneficiary, Security, Administration, Source description and service description. Under this method, we are particularly interested in the dimension "Service Description" that contains the following attributes:

- An "identifier" which ensures the uniqueness of service
- A "Title" to name the service
- A field "historical" to learn about the date of activation and deactivation of the service
- A "Type" field that provides information on the membership service sector: tourism, health, customs
- A field "End Service" that specifies whether the service is available in its latest version
- A set of Keywords that accelerate the search process of this service
- A field "government strategy" that gives information on the strategy of the government to establish this service

We represent each service S_i as follows:

$$S_i = \left(tit_i, typ_i SKW_i \right)$$

where, tit_i is the title of the service, typ_i represents the type of the service and SKW_i represents the set of keywords associated to the service S_i.

Terminology related to the field of e-gov is very rich and varied. To better manage this wealth and better guide the user to have easy access to relevant services, we propose to use a semantic layer in the form of a domain ontology.

The ontology provides a common vocabulary of e-gov domain. It defines the meaning of concepts and relations between them. We note that the five types of components which formalize the knowledge in ontologies are concepts (or classes), relationships (or properties), functions, axioms (or rules) and instances (or individuals).

Our method of evaluating the relevant services: This method contains several steps and representations. It is based on calculating the weight of each service. The best services are those with the highest weights.
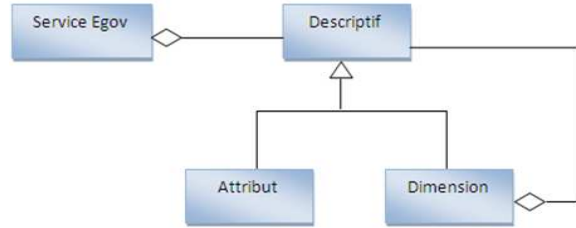

Fig. 1: The meta-model of the descriptive basis of e-government services

The queries sent by users can be represented by the set $Q = \{q_1,...,q_n\}$ and the users are represented by the set $U = \{u_1,...,u_m\}$. We note that several queries can be associated with the same user u_i. The method of evaluating the services relevance is based primarily on the treatment of concepts that are contained in the query. Therefore, the choice of mathematical representation of the query is the important element in this method. Indeed, better representation facilitates greatly the treatment of these concepts and the understanding of the method. In this context, we note t the transformation function that transforms a query of any q_i of Q $\left(q_i \in Q, \forall_i \in \{1,...,n\} \right)$ to a vector of terms VT_{qi}:

$$t : Q \rightarrow T^{ri}$$
$$q_i \rightarrow VT_{qi}$$

where, T is the set of terms (or the space of terms) and ri is the dimension of the vector $VT_{qi} \left(din T^{ri} = ri \right)$

Queries entered by users do not always contain only relevant terms. Therefore, the queries transformation provides vectors containing both relevant and unnecessary terms. In order to resolve this problem by keeping only the relevant terms, we apply a filter on the query qi. We represent this filtering mechanism as a projection function p of terms' space on another terms' space:

$$p : T^{ri} \rightarrow T^{rri}$$
$$VT_{qi} \rightarrow VTF_{qi}$$

where, VTF_{qi} represents the filtered vector and rri represents the number of the relevant terms of $VTF_{qi} \left(x_1, x_2,..., x_j,..., x_{rri} \right)$ In case where VTF_{qi} doesn't contain any unnecessary term, ri = rri.

Certainly, the filtering step of a query q_i allowed us to keep only the relevant terms, but in most cases, these terms do not fit the real user's needs. They don't correspond to those found in the vocabulary controlled

both by the service suppliers and experts in the field. In this case, the enrichment of the query is necessary by adding the various concepts related to all components x_j of the vector VTF_{qi} for $j \in \{1, \dots rri\}$.

Let be x_j a component of the vector VTF_{qi}, O_D is the domain ontology and SC_{xj} the set of the ontology's concepts which are linked to the component x_j. If x_j doesn't belong to SC_{xj} then:

$$SC_{xj} = \{x_j\} \bigcup SC_{xj} = \bigcup_{i=1}^{n_j} C_{i,j}$$

where, x_j the cardinality of the set SC_{xj} ($card(SC_{xj}) = n_j$).

We concatenate the various concepts of SC_{xj} and we apply the function t to SC_{xj}. It transforms SC_{xj} to a vector VSC_{x_j} $\left(t(SC_{x_j}) = VSC_{x_j} = (c_1^j, c_2^j, \dots, c_{n_j}^j)\right)$. The union of sets of concepts SC_{xj} for $j = 1, \dots, rri$ is noted $SC = \bigcup_{j=1}^{rri} SC_{xj}$ the transformation of SC by t is a vector VSC:

$$t(SC) = VSC = \left(VSC_{x1}, \dots, VSC_{x_{mi}}\right)$$
$$= \left(c_1^1, c_2^1, \dots, c_{n_1}^1, \dots, c_1^j, c_2^j, \dots, c_{n_j}^j, \dots, c_1^{rri}, c_2^{rri}, \dots, c_{n_{mi}}^{rri}\right)$$

The dimension of VSC is: $dim\, VSC = \sum_{k=1}^{rri} dim\, VSC_{ck} = \sum_{k=1}^{rri} nk$.

In order to retrieve the services S_i associated to a giving concept, we search it by using the services characteristics: title tit, type typ_i and set of key word SKW_i. In other word, the principle of the services retrieving process is to search all associated services to all concepts of the vector VSC (components of VSC). We define this process by an unfolding function e which associates a concept or a component of a vector of concepts to a vector of services. This function is given as follows:

$$e : T \rightarrow T^{l_{j,k}}$$
$$c_k^j \rightarrow VSS_{c_k^j}$$

where, c_k^j is the k-th component of the vector VSC_{x_j} for $j \in \{1, \dots, rri\}$, $VSS_{c_k^j}$ is the vector of the services associated with c_k^j and $l_{j,k}$ is the dimension of $T^{l_{j,k}}$ (number of services associated with c_k^j).

The set of the services associated with the vector of concepts VSC_{x_j} is a vector given as follows (for $j \in \{1, \dots, rri\}$):

$$e\left(VSC_{x_j}\right) = \left(e(c_1^j), \dots, e(c_{n_j}^j)\right) = VSSc_1^j, \dots, VSSc_{n_j}^j$$

The dimension of the vector $e\left(VSC_{x_j}\right)$ is: $dim\, e\left(VSC_{x_j}\right) = \sum_{k=1}^{n_j} dim\, VS_{c_k} = \sum_{k=1}^{n_j} l_{j,k}$.

All services VSS associated with the vector VSC, is the union of the all services related to all concepts of the VSC. All VSS is represented in a form of a vector is given as follows:

$$VSS = e(SC) = \left(e\left(VSC_{x_1}\right), e\left(VSC_{x_2}\right), \dots, e\left(VSC_{x_{mi}}\right)\right)$$
$$= \left(e(c_1^1), \dots, e(c_{n_1}^1); e(c_1^2), \dots, e(c_{n_2}^2); \dots \dots; e(c_1^{rri}), \dots, e(c_{n_{mi}}^{rri})\right)$$

where, the dimension of the VSS is:

$$dim\, VSS = \sum_{p=1}^{rri} dim\, e\left(VSC_{x_p}\right) = \sum_{p=1}^{rri}\left(\sum_{k=1}^{n_p} l_{p,k}\right)$$

During a search process, the recovered services are not distinct and some may have a very high score. This is due to several reasons such as:

- The terms entered in a query may have some similarity
- One term has several concepts in the ontology during the enrichment phase
- Only one service is identified by several words key
- We propose the following algorithm to extract the set of distinct services, presented by the array SDS, from the vector VSS and their occurrence presented by the array ODS.

```
Array VSS []: STRING;
//Set of distinct services
Array SDS []: STRING;
// Occurrence of distinct services
Array ODS []: INT;
VARIABLE j,k: INT;
// Test if a term exists in SDS
VARIABLE inSDS: BOOLEAN;
// VSS's dimension supposed already calculated
VARIABLE dimVSS: INT;
j<-0;
FOR i FROM 1 TO dimVSS DO
  inSDS <- FALSE;
  k = 1;
  WHILE (inSDS == false AND k<=dimSDS)
  DO
  BEGIN
    IF (VSS[i] == SDS[k] AND k <= dimSDS)
    THEN
```
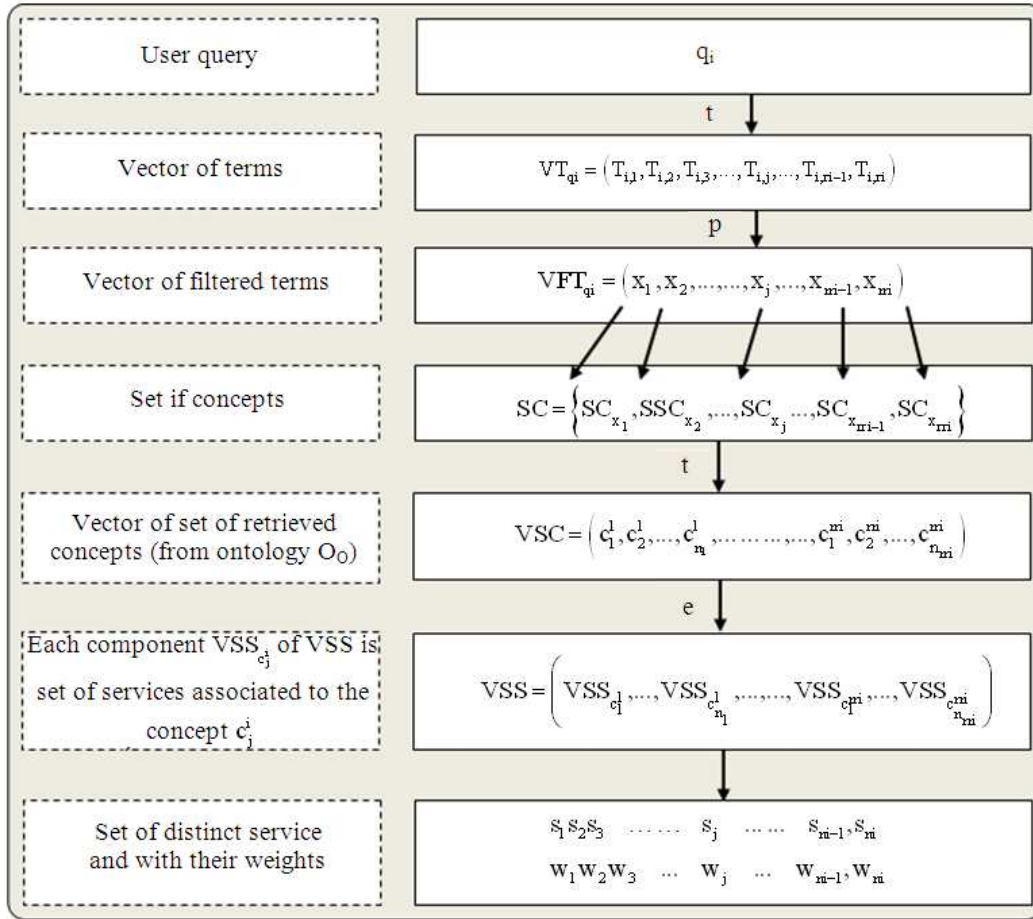
Fig. 2: Process of relevance computation

<div style="column-count:2">

```
 inSDS <- TRUE;
 ELSE
  k<-k+1;
 END IF
 ENDWHILE
 IF (inSDS == FALSE) THEN BEGIN
  j<-j+1;
  SDS[j] <-VSS[i];
  ODS[j] <-1;
  END
 ELSE
  ODS[k] = ODS[k]+1;
 ENDIF
END FOR
```

The weights of the distinct services are noted by PSD and they are calculated from their occurrences (ODS) and the dimension filled as follows:

```
Array PSD []: REAL;
VARIABLE dimVSS, dimSDS: INT;
FOR i FROM 1 TO dimSDS DO
```

```
  PSD[i] <- OSD[i]/dimVSS
END FOR
```

The presentation of the results depends on the services' relevance and consequently the weight of these distinct services. Indeed, the relevant services have a higher weight. To classify the different services given in SSD, we relied on the sort of weights:

```
VARIABLE maxIndex: INT;
VARIABLE dimPDS: INT;
VARIABLE tempReal: REAL;
VARIABLE tempString: STRING;
Array PDS []: REAL;
FOR i FROM 1 TO dimPDS-1 DO
 maxIndex <- i;
 max <- PDS[i]
 FOR j FROM i+1 TO dimPDS DO
 IF PDS[j]> max THEN BEGIN
 maxIndex <- j;
 max <- PDS[j];
```

</div>

```
END
END IF
END FOR
IF (maxIndex > i) THEN BEGIN
tempReal <- change PDS[i];
PDS[i] <- PDS [maxIndex];
PDS [maxIndex] <- tempReal;
tempString <- change SDS[i];
SDS[i] <- SDS [maxIndex];
SDS [maxIndex] <- tempString;
END
 END IF
END FOR
```

To get results as a set of distinct services and their weight, we have used several notations and steps. Fig. 2 represents these steps and notations. It greatly facilitates the understanding of our proposal and its principle without worrying about technical details.

RESULTS

We will measure the effectiveness and efficiency of the method of the evaluation of the services' pertinence proposed in this study. To perform this validation, we have integrated this method into search system proposed by Ouchetto *et al*. (2011). This system integrates several major components, but for this validation, only the domain ontology and the basic descriptive are considered and other components are not taken into account. The implementation of this method has been performed by Java language and the implementation of the system by JEE technology.

To measure the contribution and effectiveness of this method, we compare the results of services obtained by this search system integrating this method with those obtained by a direct search. We note that the direct search does not integrate any method of evaluation the services' pertinence.

We consider a sample query q_1 which contains three terms $q1 = (t1, t2, t3)$. By submitting this query to the search system, we obtain 31 different services from the descriptive basis.

We note this set of services by $S = (S_1, S_2, S_3, S_4, S_5,...,..., S_{30}, S_{31})$ where S_i is the i-th service for $I = 1,..,31$.

By using the proposed method, the weight of each service is calculated and the obtained results are presented in the following Table 1.

We note that the 31 services obtained are presented on four pages; the display option chosen is 10 services per page. To test the obtained results, this platform was made available to 20 users to give their judgment on the appropriateness of all obtained services by the search system. Their judgment is given in both cases: with the method of the pertinence's evaluation and by a direct search.

Table 1: Represents the weight of each service

Service	S1	S2	S3	S4	S5	S6	....	S31
Weight of each service	0.134	0.105	0.096	0.077	0.067	0.048	....	0.008

Table 2: Represents the relevant services existing in each page

Page	P1 (%)	P2 (%)	P3 (%)	P4 (%)
With method	69.7	21.5	7.4	1.4
Without method	17.3	35.2	19.3	28.2

The judgment was given by users in different cases. Indeed, each user gives, on the one hand, the number of relevant services among the all services and on the other hand, the number of the relevant services per page in both cases. The number of relevant services for all users is presented as a percentage. Among all rendered services, 24% are judged relevant (7.44 among 31 services). The following table shows the number of relevant services per page in both cases.

DISCUSSION

It is not difficult to see from the results in Table 2 that the newly discovered ranking method performed very well in the search context. The relevant services in the first page and the order of relevance decreases with the pages. It gives a satisfaction for the users. In the case of a direct search, the results of relevance remain highly uncertain, random and not subject to any rules.

The domain ontology is a very important component in both the search system and in our proposed method. Having a well designed ontology and rich allows greatly to improve the performance of our method and have high accuracy of the obtained results. As perspective of the present study, we intend to integrate this method in the other type of retrieval system and other type of data basis.

CONCLUSION

We have proposed a new method for automatic ranking of retrieval systems. This approach integrates a search system which incorporates a semantic layer and descriptive base of services as the crucial elements in the retrieving services process.

To experiment the proposed method, we use the e-gov domain. In the other way, our descriptive base contains the e-gov services. The obtained results were compared with the direct search without using any method of ranking. The experiments show a high level of satisfaction of this method by improving the quality of the presentation and the relevant services are presented in the first page and the order of relevance decreases with the pages.

In the perspective of the present sutdy, we intend to integrate this method in the other type of retrieval system and other type of data basis.

REFERENCES

Bhogal, J., A. Macfarlane and P. Smith, 2007. A review of ontology based query expansion. Inform. Proc. Manage., 43: 866-886. DOI: 10.1016/j.ipm.2006.09.003

Gauch, S., J. Chaffee and A. Pretschner, 2003. Ontology-based personalized search and browsing. Web Intell. Agent Syst., 1: 219-234.

Kowalski, G.J. and M.T. Maybury, 2000. Information Storage and Retrieval Systems: Theory and Implementation. 2nd Edn., Springer, Boston, ISBN-10: 0792379241, pp: 336.

Liu, F., C. Yu and W. Meng, 2004. Personalized web search for improving retrieval effectiveness. IEEE Trans. Know. Data Eng., 16: 28-40. DOI: 10.1109/TKDE.2004.1264820

Mianowska, B. and N.T. Nguyen, 2011. A Method of User Modeling and Relevance Simulation in Document Retrieval Systems. Proceedings of the 5th KES International Conference on Agent and Multi-Agent Systems: Technologies and Applications, (AMSTA' 11), Springer-Verlag Berlin, Heidelberg, pp: 138-147.

Middleton, S.E., N.R. Shadbolt and D.C.D. Roure, 2004. Ontological user profiling in recommender systems. ACM Trans. Inform. Syst., 22: 54-88. DOI : 10.1145/963770.963773

Ouchetto, H., O. Ouchetto and O. Roudies, 2012. Ontology-oriented e-gov services retrieval. Int. J. Comput. Sci. Issues.

Ouchetto, H., O. Roudies and O. Ouchetto, 2011. A personalised access to e-gov services: A user-oriented approach. Int. J. Sci. Adv. Technol., 1: 101-106.

Rijsbergen, C.J.V., 1979. Information Retrieval. 2nd Edn., Butterworth, London, ISBN-10: 0408709294, pp: 208.

Robertson, S. and I. Soboroff, 2002. The TREC 2002 filtering track report. Proceeding of The 11th Test Retrieval Conference, NIST Special Publication, pp: 500-251.

Robertson, S.E., M.E. Maron and W.S. Cooper, 1982. Probability of relevance: A unification of two competing models for document retrieval. Inform. Technol.: Res. Dev., 1: 1-21.

Sugiyama, K., K. Hatano and M. Yoshikawa, 2004. Adaptive web search based on user profile constructed without any effort from users. Proceedings of the 13th International Conference on World Wide Web, (WWW' 04), ACM New York, NY, USA, pp: 675-684. DOI: 10.1145/988672.988764

Wallis, P. and J.A Thom, 1996. Relevance judgements for assessing recall. Inform. Proc. Manage., 32: 273-286. DOI: 10.1016/0306-4573(95)00061-5

Yu-Ming, Z., X. Zhi-ming, W. Xiao-long and Z. Kun-Peng, 2008. Document relevance calculation based on lexical cohesion. J. Electron. Inform. Technol. 30: 2512-2515. DOI: 10.3724/SP.J.1146.2007.00493