A geo-social user profile for a personalized information retrieval

RAFA Tahar
Department of Mathematics and Computer Science.
University of MEDEA, Algeria.
(rafa.tahar@gmail.com, rafa.tahar@univ-medea.dz)

KECHID Samir
LRIA Laboratory, Department of Computer Science
USTHB, Algiers, Algeria.
(Kechidsam@yahoo.fr)

ABSTRACT
Several works in user-centered personalized information retrieval treat independently the user's mobile and social contexts, and they develop and exploit separately the social and situational user profiles to improve the information access process. We propose, in this paper, a personalized information retrieval approach combining the social and situational user profiles to improve the search results relevance. We intend to improve two important phases of the research process, (i) user query expansion and (ii) adaptation of search results to the user profile.

CCS Concepts
Information systems →Information retrieval →Users and interactive retrieval →Personalization

Keywords
Personalized information retrieval; social profile; situational profile; profiles combination.

1. INTRODUCTION
The personalized information retrieval aims to introduce the user dimension to improve the information access process [1], through the construction of a structure to storing data about the user and his interests. This structure, called User Profile [2], is used to adapt the results returned by the search engines to the user interest. The recent works of information retrieval that focus on the user modeling exhibit two basic types of user profiles:

(i) *Social profile* [3], [4], [5] and [6]: This profile is pushed by the expansion use of social networks. It considers the user community as friends, neighbors, colleagues, etc [7]. This profile is built from search history, tags and social relationships with other social network users, etc. The works that develop this profile support the hypothesis that the user social information (annotations, comments, sharing, and relations with other users of the social website, etc.) Can help to infer the user relevance and identify their information needs.

(ii) *Situational profile* [8], [9] and [10]: it is driven by the emergence of smart mobile devices which remotely enable search and access to information. In this profile, the information importance is related to the user geographical location in the search time [7]. This profile is built from information about the user locations and the search history related to these locations.

The works that develop this profile support the hypothesis that the user information need can be related to user current location. And therefore, the information about the user location such as name, type, GPS coordinates, etc, Can help to identify this need.

We noticed, from our literature search, that the user-centered information retrieval approaches develop and exploit independently the user's social and situational profiles. While their operating jointly, as we imagine, can better help to identify the user information needs hidden by his query, and to improve the search results relevance.

In this paper, we propose an information retrieval personalized approach based on building a user profile that combines his social and situational contexts. The basic idea of this work is that the mobile user query, often short and ambiguous, can hide behind it information needs related to the user's current geographic location, or general information needs. So, the design of a user profile regrouping, in an evolutionary framework, the user interest indicators from both social and situational contexts of previous researches, can best help to disambiguate the user query and identify the information need hidden behind it. Our goal is to use this profile to the query expansion and the adaptation of research results to the user interests.

The rest of the paper is organized as follows: section 4 is for related works, Section 5 is devoted to the presentation of our approach. This section contains two main sub sections, one to present the different components of user profile and the other to present the different phases of our search process. In Section 6 we present the evaluation of our approach. And we end with a conclusion and perspectives.

2. RELATED WORKS
Several works in the personalized information retrieval, exploit social information to improve the search results. Saoud [3] provides, in a distributed information retrieval framework, a user profile that uses social annotations "folksonomies". Bouhini [5], exploits social annotations to build the user profile, also she exploits it in the document indexing phase. Benjabeur [6] and [11] offers a social user profile based on social relationships between users. He introduced the concept of user social importance or user influence in the social network.

Other works exploit situational information to improve the information search results. Akermi [8] provides a contextual recommendation approach in a mobile environment guided by a user situational profile. This profile is based on user location and his preferences, and aims to identify the right information to

ICIME 2016, November 02-05, 2016, Istanbul, Turkey
© 2016 ACM. ISBN 978-1-4503-4761-7/16/11…$15.00
DOI: http://dx.doi.org/10.1145/3012258.3012270

recommend depending on the user current situation. Bouidghaghen [9] offers a personalized search approach based on building of a user situational profile composed by the user's research situations, and his related interests. In [10], the authors propose a mobile user profile in which the user's situations are represented by the most visited places.

In summary, as is observed in [4], [2], [12] and [7], we have not found works that combine social and situational information to improve search results.

3. PROPOSED APPROACH

For us, relevance of a user of a social website using a mobile device may be deducted in addition to his query, from: (i) the thematic aspect, represented by search history that includes the user's important documents (long visited, many clicked, printed, ...). (ii) The social aspect, represented by the history of user's social operations (annotation, tagging or sharing documents). And (iii) the situational aspect consists of information about different geographical locations from which the user has made information searches.

In objective to identify and to group data that can contribute to the description of user relevance, we will build a user *geo-social profile* that includes data characterizing the three aspects already mentioned.

3.1 User Geo-Social Profile Modeling

The user geo-social profile that we offer, aims the description of the user relevance by combining data characterizing the previous three aspects: thematic, social and situational. This geo-social profile consists of three (03) bases: (i) base of research situations, (ii) base of relevant terms and tags related to the location, and (iii) base of relevant terms and tags qualified of general order.

3.1.1 The Research Situations Base (RS_B)

It includes all research situations related to different locations from which the user has made information searches. A research situation includes all query issued in one location. Each research situation (RS) consists of three components:

(i) A situational component:

It includes all information about the user location. A location is represented by several parts: (place name, city, region, country). These parts are linked by a membership relation in this order: (place Є city Є region Є country). And each place is identified by a name and a type.

The location parts are derived by comparing the GPS coordinates of the user mobile device to a geographic database inputs.

(ii) A thematic component:

It includes:

- Search History related to the research situation. This history is represented by all documents considered relevant by the user (long visited, printed, shared, annotated, tagged, etc.)
- A set of *K* more weighted terms from the research situation history.

The set of the more weighted terms form the interests center of the research situation, it is represented by a vector noted:

$$Vt=\{ (t1, wt1), (t2,wt2), ..., (ti, wti), ..., (tk, wtk) \} \quad (1)$$

Where:

- t_i: term number "*i*" in interests center.
- wt_i: represents the overall weight of t_i relative to the research situation history considered as a documents collection.

The weight of a term *t* in an RS history is calculated using the formula *tf * idf* of the vector model [13]:

$$wt = (0.5 + 0.5 * tf / Max\ tf) * \log (N / n) \quad (2)$$

With:

- *tf*: total frequency of term *t* in the document history.
- *Max tf*: the greater frequency of terms in history documents.
- *N*: number of documents of the search history.
- *n*: number of documents of search history containing *t* .

(iii) A social component:

Represented by the set of *K'* of user's most used tags to annotate relevant documents of the research situation. In this base we consider only the significant tags (terms or concepts). This set of tags is represented by a vector denoted:

$$Vg = \{ (g_1,wg_1), (g_2,wg_2),..., (g_i,wg_i), ... (g_{k'},wg_{k'})\} \quad (3)$$

Where:

- g_i: tag number "i" in the vector.
- wg_i: represents the overall weight of g_i relative to the current research situation history.

The weight of a tag g is calculated according to the formula *tf *idf* of the vector model. It is given by the following formula:

$$wg = (0.5 + 0.5 * gf / Max\ gf) * \log (N / n) \quad (4)$$

With:

- *gf*: total frequency of *g* in documents of search history .
- *Max gf*: largest total frequency of tags in current RS history.
- *N*: number of documents in search history.
- *n*: number of documents annotated by *g* in search history.

Tag-term Coexistence

A boolean coexistence relationship is defined between each frequent tag g and each relevant term t of the interest center of the research situation. It is denoted *Coexist(g,t)*, and is defined as:

$$\textbf{Coexist(g,t)=} \begin{cases} \textbf{1} & \text{if there is at least one relevant document containing } \textbf{t} \text{ and annotated by } \textbf{g}. \\ \textbf{0} & \text{if there is no relevant document containing } \textbf{t} \text{ and annotated by } \textbf{g}. \end{cases}$$

3.1.2 Base of Relevant Terms and Tags Related to the Location (B_LOC)

This base (*B_LOC*) is supplied from *RS_B*. It gathers all terms and tags that are relevant for user, and are linked to the location. The link of a term (respectively a tag) to one location of a research situation is expressed by a *degree of link* calculated with each location part (place name, place type, city, region, country). This degree is calculated at two levels: (i) a *simple link* to one location part of the current research situation, and (ii) a *global link* calculated using simple links in all research situations that their locations include the location part concerned.

The simple link to location of terms and tags

In a given research situation, the simple link degree of a term t (respectively a tag g) to one of its location parts (loc_i); noted *L (t, loc_i)* (*L (g, loc_i)* respectively) is calculated as follows:

- $L(t, loc_i) = |D(t, loc_i)| / |D(t)|$ (5)
- $L(g, loc_i) = |D(g, loc_i)| / |D(g)|$ (6)

Where:

- loc_i: one location part of the current research situation of (place name , place type, city, region, country).
- $|D(t, loc_i)|$: number of documents in situation history that contain t and referring to the location part concerned loc_i.
- $|D(t)|$: Number of documents in research situation history containing the term t.
- $|D(g, loc_i)|$: number of documents in situation history that are annotated by g and referring to loc_i.
- $|D(g)|$: Number of documents in situation history annotated by the tag g.

The global link to location of terms and tags

Figure 1 presents the modeling of the global link between a term (or tag) and the different location parts.

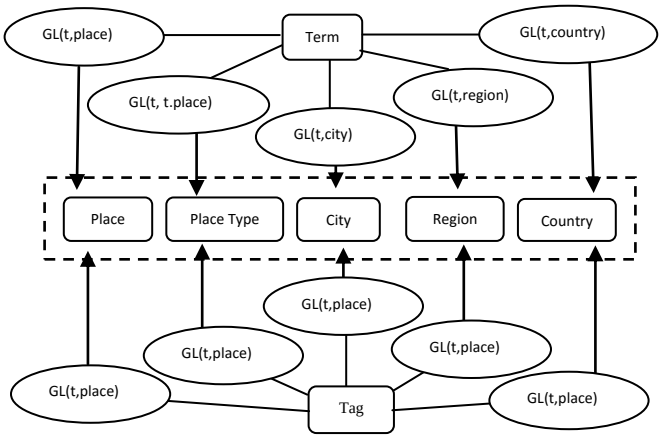

Figure 1. Modeling of the global link to location

After each instantiation of a new research situation (or update an existing situation), we calculate the global link of all interests center terms, and all frequent tags with each location part of this research situation. The global link of a term (or tag) at a location part, is calculated from all other research situations of geo-social profile. A global link of a term *t* to a location part (loc_i) of a given research situation, noted $GL(t, loc_i)$, is calculated as follows:

$$GL(t,loc_i) = (|RS(t, loc_i)|/|RS|^2) * \sum_{j \in |RS(t, loc\ i)|} (wt_j * L_j(t, loc_i)) \quad (7)$$

With:

- $|RS(t, loc_i)|$: Number of RS including loc_i, and in which the simple link degree ($L(t,loc_i) >= 0.25$).
- $|RS|$: Total number of research situations in the user profile.
- wt_j: weight of term t in the search history of the research situation RS_j by (Formula 2).
- $L_j(t, loc_i)$: simple link degree of the term t to loc_i of RS_j (Formula 5).

Similarly we calculate the global link of a tag g to a location part (loc_i), denoted $GL(g, loc_i)$ with respect to all user profile research situations as follows:

$$GL(g,loc_i) = (|RS(g, loc_i)|/|RS|^2) * \sum_{j \in |RS(g, loc\ i)|} (wg_j * L_j(g, loc_i)) \quad (8)$$

Where wg_j is calculated by (formula 4), and $L_j(g, loc_i)$ by (formula 6).

5.1.3. The base of general relevant terms and tags (B_GNL)

This base is used to store the terms and tags relevant for the user independently of any location.

• Degree of No Link to location

In a research situation, the degree of "No Link" of a term *t* (a tag *g*, respectively) to the location, noted *NL (t, loc),* (and *NL (g, loc)*, respectively), is calculated as follows:

$$NL(t, loc) = |D(t, N\ loc)| / |D(t)| \quad (9)$$

$$NL(g, loc) = |D(g, N\ loc)| / |D(g)| \quad (10)$$

With:

- $|D(t, N\ loc)|$: Number of Documents of the RS history containing t and not referring to any location part.
- $|D(t)|$: Number of Documents of the RS history containing the term t.
- $|D(g, N\ loc)|$: Number of Documents of the RS history annotated by g and not referring to any location part.
- $|D(g)|$: Number of Documents of the RS history annotated by the tag g.

Degree of general order qualification of terms and tags

For us, more a term (or tag) is relevant in many research situations, and independent of their locations, most it is qualified to general order. The degree of general order qualification of a term t, denoted GNL(t) is calculated after the instantiation of each new research situation, from all other research situations of geo-social profile as following :

$$GNL(t) = (|RS(t, Nloc)|/|RS|^2) * \sum_{i \in |RS(t, Nloc)|} (wt_i * NL_i(t, loc)) \quad (11)$$

With:

- $|RS(t, Nloc)|$: Number of research situations in which ($NL(t, loc) >= 0.25$).
- $|SR|$: Total number of research situations in the user profile.
- wt_i: weight of term t in the history of RS_i calculated using (Formula 2).
- $NL_i(t, loc)$: degree of no link of the term t to the location of RS_i (formula 9).

Similarly we define the degree GNL (g) for a tag g, as follows:

$$GNL(g) = (|RS(g, Nloc)|/|RS|^2) \sum_{i \in |RS(g, Nloc)|} (wg_i * NL_i(g, loc)) \quad (12)$$

With wg_i is calculated using (formula 4) and $NL_i(g,loc)$ using (formula 10).

3.2 The Search Process

Our personalized information retrieval process, as shown in Figure 2, follows the following scenario:

- The system receives the user query (Q), and recovers its location.
- The system selects the profile base (B_LOC or B_GNL) most appropriate to user query.
- The query Q is reformulated (enriched), from the selected base.
- The reformulated query Q' is sent to the classic search engine.
- The elements (terms and tags) of Q' are associated to their weights in the selected base.

- The results returned by the classic search engine will be submitted to another matching operation, to rank them according to their similarity scores to the weighted query Q'.
- The new reordered document list is transmitted to user.
- The system observes the reactions and behavior of the user to update his profile.

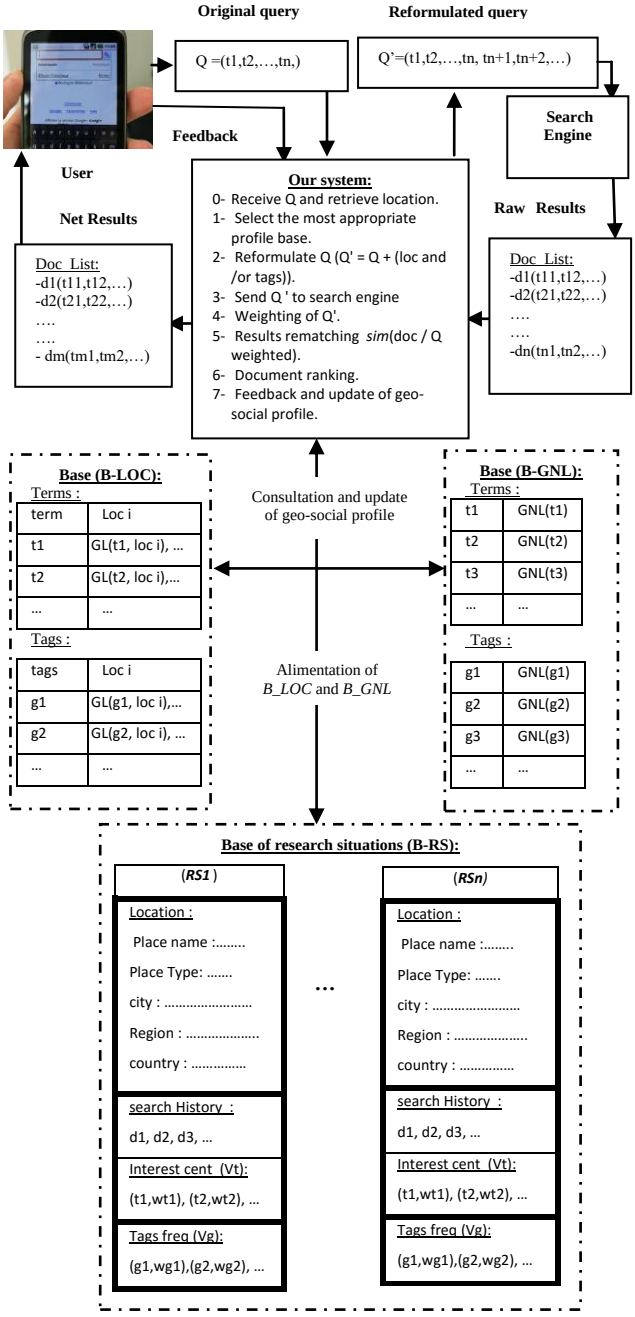

Figure 2. Modeling of the user profile and research process

3.2.1 Selection of the profile base most appropriate to user query

To select from two profile bases *B_LOC* and *B_GNL*, the most appropriate to the user query Q, our system calculates a similarity score between the query and each of the two bases as follow:

$$Sim\ (B\_LOC/Q)= \sum_{loci \in current\ location} \sum_{t \in Q} GL(t, loc_i) \quad (13)$$

$$Sim\ (B\_GNL/Q)= \sum_{t \in Q} GNL(t) \quad (14)$$

The most similar base to user query will be selected.

3.2.2 Query reformulation

With the aim of widening the search field without going further than the user interests, the user initial query (Q) will be enriched by the injection of some relevant terms and/or tags provided from the selected profile base.

- if *B_LOC* is selected, the user's initial query is reformulated to produce a new query Q', such as:

$$Q'= Q+ P+G \quad (15)$$

With:

- Q = (t1, t2, ... tn): the initial user query.
- P = {$loc_i \in$ current location, such as: *ther is* $t_q \in Q$ *with* $GL(t_q, loc_i) > 0$}: set of current location parts to which is linked at least one term of query Q.
- G = {$g_j \in B\_LOC$, such as: (*ther is* $loc_i \in P$, *with* $GL(g_j,loc_i)>0$) *and (ther is* $t_q \in Q$ *with* $coexist(t_q,g_j)=1$)}: set of tags having a coexistence relationship with at least one term of Q, and that are linked at least to one element of P. if the number of tags concerned is important, we take only the tags having global link exceeding the average of tag global links in P.

- if *B_GNL* is selected:

$$Q'= Q+ G' \quad (16)$$

With:

- G' = {$g_j \in B\_GNL$, such as: *there is* $t_q \in Q$ *with* $coexist(t_q,g_j)=1$} set of tags having a coexistence relationship with at least one term of Q. If the number of involved tags is important, we retain only the tags having a degree of general order qualification above average in the tags base of *B_GNL*.

3.2.3 The relevance Recalculation and the document ranking

In this step we proceed to redo the matching operation of results returned by the search engine to the reformulated query Q', taking into account the weight of terms and tags. Each term or tag of Q' appearing in the selected profile base, is associated with its weight from this base. Other elements of Q' such as initial terms that do not appear in the selected profile base, and added location parts, If it's happened, are not concerned by this step of weighting and matching.

The matching operation result is a similarity score for each document returned to the query Q' weighted. And according to these scores, the documents will be reclassified in descending order, thus defining the final list of the most relevant documents to transmit to user.

To lighten this re-matching task, we will consider only the list of top *N* documents returned by the search engine. The similarity of a document *d* to the weighted query Q', noted *Sim (d, Q')* is calculated as follows:

- If *B_LOC* is selected:

$$Sim(d,Q')=\sum_{t \in Q' \cap B\_LOC}((tf_d/|d|)*\sum_{loci \in P} LG(t,loc_i))* log(N/rang(d)) \quad (17)$$

With:

- *t*: an element of reformulated query Q' (term or tag) classified in B_LOC.
- tf_d: frequency of element t in the document d.

- | d | : Size document.
- Loc_i: element of the set P of the current location parts present in Q'.
- *GL (t, loc_i)*: global link of the element t to loc_i.
- *N*: number of top returned documents.
- *Rank (d)*: grading of document *d* in the N top documents.

- If *B_GNL* is selected:

$$Sim(d,Q') = \sum_{t \in Q' et\ B\_GNL} ((tf_d/|d|) * GNL(t)) * log\ (N/rang(d)) \quad (18)$$

With:

- *t*: an element of reformulated query Q'(term or tag) classified in *B_GNL*.
- *GNL (t)*: degree of general order qualification of t.

Update of User Profile

The update task should ensure the alimentation and the optimization of geo-social profile. We aliment the user profile by instantiation of a new research situation or updating an existing research situation. And we optimize it by removing of any research situation whose date of its last update exceeds one year.

4. THE APPROACH EVALUATION

In the absence of an evaluation framework to the mobile context, We will evaluate our approach with a simulation in which we have differently constructed geo-social profiles for 15 users from different professional fields (teachers, students, doctors, athletes, tourists). Each user sends 5 queries from different locations. In sum 65 queries and 12 locations (some queries and some locations are common). We have used the search engine *Google* for research, and our evaluation is to calculate the precision factor at the 4, 8 and 12 first documents returned by Google for each of 3 following scenarios: (A) query sent without reformulation, (B) query reformulated only by the location information, (C) query reformulated using our geo-social profile. The following table (table 1) presents the simulation results.

Table 1. Simulation results.

precision	Scenario (A)	Scenario (B)	Scenario (C)	Rate (C/A)	Rate (C/B)
P @4	0.40	0.37	0.45	5 %	8 %
P @8	0.48	0.43	0.52	4 %	9 %
P@12	0.52	0.49	0.58	6 %	9 %
Average	0.466	0.356	0.516	5 %	**4.66**

5. Discussion

We note that the results without reformulation are better than those with a reformulation only with situational information. Indeed, this amounts to fact that the majority of queries are formed so that they can be seen locatable by any search engine, while according to user profiles, they are not locatable. Simulation details show that our results are good in two cases: (i) very short query and very rich profile, and (ii) in case of query which can be seen localizable by search engines while it is not according to the user interests. Our results are limited in case of new profile or in case of an explicit localizable query.

6. CONCLUSION AND FUTURE WORKS

In this paper we have proposed a new personalized information retrieval approach, based on the definition of a geo-social profile that combines the interest indicators provided from both social and situational contexts of the user's previous researches. This approach is proposed in objective to improve the search results relevance by better adapting to the user interests. The simulation results, especially for the query reformulation phase, show that over the user profile is richer, the query is better reformulated and search results are more relevant. This work opens way to future works such as adding a base of terms related to events and the introduction of semantics in user profile.

7. REFERENCES

[1] Kechid, S. 2009. *integration du modèle utilisateur dans un système de recherche d'information distribuée*. doctoral thesis. USTHB, algeria, 2009.

[2] Farida, A. and Rachid A. O. 2012. Modélisation *d'évolution de profil utilisateur en recherche d'information personnalisée*. CORIA 2012. 83-97.

[3] Zakaria, S., Samir, K., and Radia, A., 2014. *Exploring Folksonomy Structure for Personalizing the Result Merging Process in Distributed Information Retrieval*. in springer international publishing, Switzerland 2014. 42-50.

[4] Bilel, M., Lynda, T. and Sadok, B. Y., 2014. *Prise en compte des préférences des utilisateurs pour l'estimation de la pertinence multidimensionnelle d'un Document.* INFORSID 2014. France. 42.

[5] Chahrazed, B., Mathias, G. and Christine, L. 2013. *Modèle de Recherche d'Information Sociale Centré Utilisateur*. (EGC'2013). France. Journal of new IT, <ujm-00869337>. 275-286.

[6] Lamjed, B. J., Lynda, T. and Mohand, B., 2011. *Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter*. 2nd Conference on Models and Analysis of Networks: Approaches Mathematics and Computer Science, MARAMI 2011, Grenoble.

[7] Lynda, T. L., 2008. *De la recherche d'information orientée système à la recherche d'information orientée contexte: Verrous, contributions et perspectives*. HDR memory, Univ Paul Sabatier - Toulouse III, 2008.

[8] Imen, A., Mohand, B. and Rim, F., 2015. *Une approche de recommandation proactive dans un environnement mobile*. INFORSID 2015, France. 301-316.

[9] Ourdia, B., 2011. *Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes*. Phd thesis, Univ Toulouse III - Paul Sabatier, France 2011.

[10] Bila, N., Cao, J., Dinoff, R., Ho T., Hull R., Kumar B., Santos P. 2008. *Mobile User Profile Acquisition Through Network Observables and Explicit User Queries*. 9th Int'l conference on Mobile Data Management, 98-107, 2008.

[11] Lamjed, B. J., Lynda, T. and Mohand, B., 2010. *A social model for Literature Access: Towards a weighted social network of authors. International Conference on Adaptivity*, Personalization and Fusion of Heterogeneous Information (RIAO 2010), France.

[12] Abdelkrim, B., Mohamed, K. K. and Bich, L. D., 2010. *PRESY: A Context Based Query Reformulation Tool for Information Retrieval on the Web*. In Journal of Computer Science 6 (4): 470-477, 2010. ISSN: 1549-3636.

[13] Salton, G. and Buckley, C., 1988. *Term Weighting Approaches in Automatic Text Retrival*. Information Processing and Management. 513-523.