An Event-based 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, Alegria.
(Kechidsam@yahoo.fr, skechid@usthb.dz)

Abstract— **The user of a social website is often interested by information about events that have interested other users, especially his friends, and that appear remarkably in their researches, exchanges and sharing. In this paper, for a better description of the user and his interests, we propose: (i) a geo-social user profile that takes into account the event aspect of user information needs, (ii) a search process integrating the user profile for a better personalization and adaptation of the search results.**

Keywords—Personalized information retrieval, geo-social user profile, profile combination, event extraction.

I. Introduction

The objective of personalized information retrieval PIR is to enable the information retrieval systems to return, in response to user query, the best relevant results to his real information needs [1]. The personalization of the information retrieval is based on taking into account, a set of data about the user and his interests. These data, which can be explicitly expressed by the user or implicitly extracted from his previous research history, cover several contextual aspects: thematic, social, situational ... etc [2], [3]. Another very important aspect for describing the user interests is the event aspect. Indeed, in some cases, the real information need of user query is prompted by an important event that occupies public opinion on the social network. Thus, taking into account events and representing them in the user profile may help to improve the search results relevance [4].

This paper is a continuation of our previous work [2], whose basic assumption is that the user query can be prompted by an information need related to his location or his social context. In this work we will expand the coverage to take into account the event dimension of the information needs hidden behind the user query. Our new basic assumption is: The user may be interested by current, recently-passed or future events, that attract the attention of a large part of social website community, and he may be interested by past events having a periodicity of reproduction or a periodicity of celebration. In this work we will: (i) propose an event model to facilitate its identification and extraction from text, (ii) extend the geo-social user profile proposed in [2] to enable it to take into account events having interested the user and his close community on the social network, and (iii) enrich the search process with another formula-set to calculate the different weights and similarities about events and their associated terms and tags.

The document is organized as follows: after the introduction, section 2 is reserved to related works, section 3 provides details about the event model, in section 4 we present the user profile followed by the search process in Section 5, an experimental simulation is presented in Section 6, and we finish with a conclusion and perspectives.

II. Related Works

In the literature of information retrieval, temporal information often taken into consideration is that concerning the document (creation, update …) [4], or that concerning the search time (the query sent time) [5]. The objective is to be able to classify documents according to this temporal information. Nevertheless, very few approaches are interested by temporal and event aspect of the document contents [4]. In [6], an information retrieval approach is proposed which combines the temporal information on the document and that of its content. The authors of [7] propose a semantic-temporal system for learning the extraction of events and named entities from text. In [8], the user query is considered as bearer of both thematic and temporal sense. The work presented in [4] is somewhat close to our vision. It proposes a system for the extraction and indexing of events and the retrieving of their related subjects from texts. We note that we have not found an information retrieval approach that focuses on events in a context of personalization.

III. Event Modeling

A. Event definition

The notion of event does not lead to a strict, precise and consensual definition [9]. Indeed, the bibliographic research we have done, allows us to observe a very wide variety of event definitions. These definitions differ according to the field of study (history, philosophy, linguistics ...); nevertheless they contain a common characteristic of event "*anchored in time*" [10].

In our case, we adopt the definition proposed in [9]: "*the events are finite spatiotemporal entities*". And that proposed in [11]: "*an event E is the combination of three components E(S,SP,I): a semantic property S, a temporal interval I, and a spatial entity SP*".

B. *Events Classification*

The event models proposed by the different approaches of event extraction allow extracting from the texts anything that can be judged as event, without regarding to its importance for the user. This goes beyond our needs in this paper, which focuses only on events important to users, and the social website community. In our case, and to avoid the problem of diversity of classes proposed in the different works to describe an event, we will classify the event in only two classes:

- *Past Event*: concerns events having a periodicity of reproduction (the football world cup for example), or those having a periodicity of celebration (the world war for example).
- *Current Event*: concerns an event that has recently passed and still holds the community's opinion, an event in progress (arrived and not yet completed), or an event that will come very soon (planned) but it attracts attention now.

C. *Proposed Event Model*

For us, an event consists of three (03) components (Figure 1):

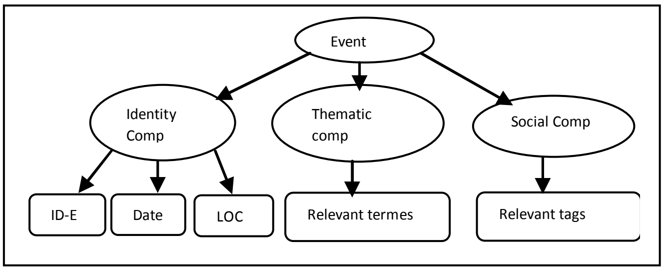

Fig. 1. Event model

- *Identity component*: Has three properties
- Event identifier: can be represented by a proper name, named entity, or even by a sentence.
- Event date: may be in a full or partial form of date.
- Event location: can be represented by the city name, region, country ... etc.

To extract this event information from the text, we can use the temporal markup language TimeML [12], particularly the attributes: <EVENT> which allows identifying the event identifier and <TIMEX3> which allows extracting and normalizing the temporal expressions related to the event.

- *Thematic component*: to store relevant terms related to the event. These terms are extracted from the user's search history.
- *Social component*: to store tags most used by the user to annotate documents referring to concerned event.

To each term and tag linked to event is associated a score expressing the degree of this link (explained in the following).

IV. MODELING OF THE PROPOSED EVENT-BASED GEO-SOCIAL USER PROFILE

The geo-social user profile proposed in [2] consists of three bases. (i) The research situations base (*B_RS*) used to store relevant search history (documents, locations, terms and tags). (ii) Base of relevant location-linked terms and tags (*B_LOC*). (iii) Base of relevant terms and tags of general order (*B_GNL*). We propose in this paper an extension of this geo-social profile to take into account the event aspect of user interests, by introducing two other bases to store past events, current events, and their linked terms and tags (Figure 2).

Note: Calculations for tags and terms are applied similarly. So, in the following, we will present only calculations for terms.

A. *Base of relevant past events and their associated terms (B_Event_P)*

This base is used to store the most frequent past events in documents of *B_RS*. For each event is associated a set of its most linked K-terms (and K'-tags).

1) *Weighting of a past event*

For us, more an event is present in many documents of different research situation (RS), more is relevant. So, the weight of a past event E, denoted WP (E), is calculated as follows:

$$WP\,(E) = (|D\,(E)|/|D|) * (|RS\,(E)|/|\,RS|) \qquad (1)$$

With:

- | D (E) |: Number of documents in B_RS containing event E.
- | D | : Total number of documents in B_RS.
- | RS (E) |: Number of RSs in B_RS containing at least one document containing event E.
- | RS | : Total number of RSs in B_RS.

2) *Weighting of terms related to a past event*

The link of a term to an event is expressed by a link degree calculated at two levels:

a) *The simple link of terms to a past event*

In a given RS, the simple link degree of a term *t* to an event *E* is a proportion of the number of documents containing *t* and referring to *E*, with respect to all relevant documents containing *t* in the RS.

This degree, denoted L(t, E), is calculated as follows:

$$L\,(t, E) = |\,D\,(t, E)\,|\,/\,(D)\,| \qquad (2)$$

With:

- |D (t, E)|: number of documents in RS's history containing the term t and referring to event E.
- |D (t)|: Number of documents in RS's history containing the term t.

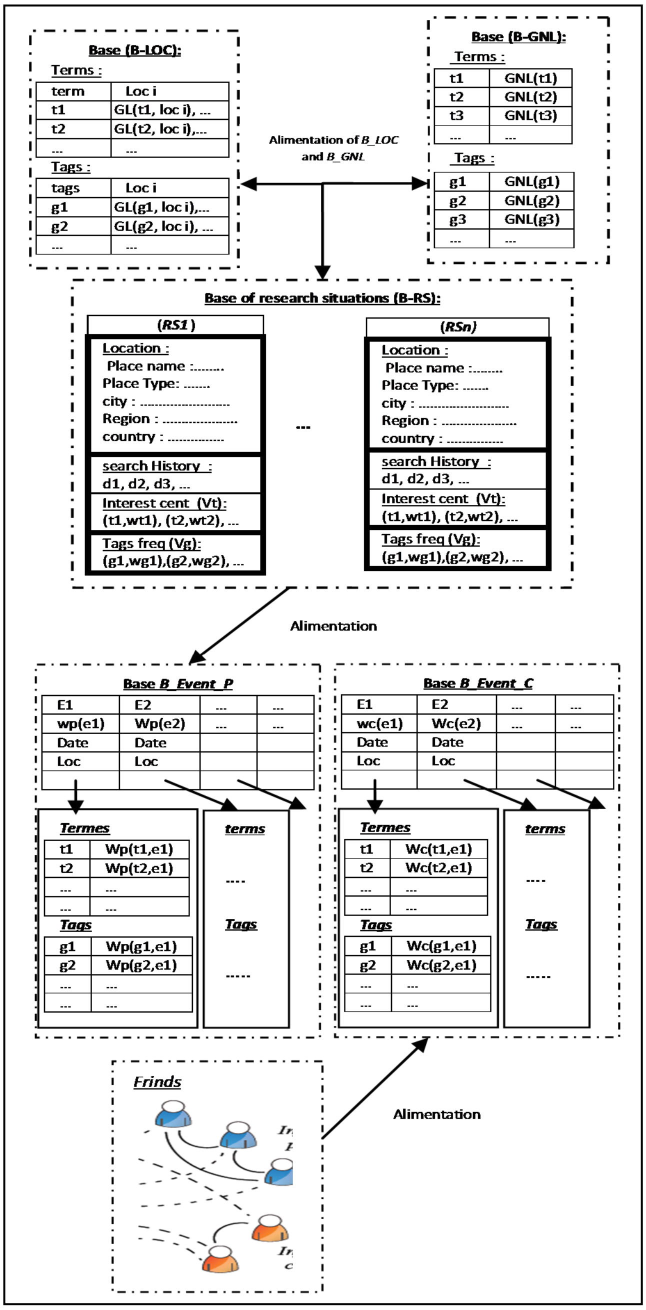

Fig 2. Event-based geo-social user profile

b) The global link of terms to a past event

A global link of a term to an event is calculated, after each instantiation of a new RS, relative to all other RS in the event-based geo-social profile. For this link are favored, terms that meet the two criteria: (i) Presence jointly with the event in many RS, and (ii) In each RS where they appear, they are relevant and they are very related to concerned event.

A global link of a term t to event E, denoted LGP (t, E), is calculated as follows:

$$LGP\ (t,E)= (|RS(t, E)|/|RS|^2)* \sum_{i \in |RS(t, E)|} (wt_i * L_i\ (t, E)) \qquad (3)$$

With:

- |RS (t, E)|: Number of RS in which the degree of simple link of the term t to the event E is important (L (t, E)> = 0.25).
- |RS|: Total number of RS in the user profile. This denominator is squared to give importance to the number of RS in which the term is related to the event relative to the total number of RS, and to normalize the formula of LGP (t, E). i.e. (0 <= LGP (t, E) <= 1).
- wt_i: weight of the term t in the history of RS_i (according to the formula *tf* * *idf* of the vector model $wt = (0.5 + 0.5 * tf / Max\ tf) * \log (N / n)$ [13]).
- L_i (t, E): degree of simple link of the term t to the event E in RS_i (formula 2).

The formulas (2 and 3) are calculated by the same principle used for location linked terms in [2].

B. Base of current events and their associated terms (B_Event_C)

News events may not appear in the user's search history. Nevertheless, they may interest him in fact that they interest his friends. To help the user about current events, we will provide to his profile a base to store the most current events in the search history of his friends. Each event is associated with a set of the most related terms and tags.

1) Friend weighting

For us, the importance of a relevant event in a friend's search history comes from the importance of the friend himself to the user. We borrow the notion of the user social importance in a social network proposed in [14], based on the number of retransmissions of its articles by the user on the social network.

So, the weight of a friend *f*, denoted W (*f*), is calculated as follows:

$$W (f) = |D\_SH (f)| / |D\_SH| \qquad (4)$$

With:

- |D_SH (f)|: Number of documents received from friend f and shared by the user.
- |D_SH|: Total number of documents shared by the user.

2) Weighting of a current event

Let F = {(f_1, wf_1), (f_2, wf_2)... (f_n, wf_n)}, the set of the user's friends , jointed by their importance degrees to him. The weight of a current event for a user, denoted WC (E), is given by:

$WC(E) = (1/|F|)*\Sigma_{i=1..|F|}(wp_i(E)*w(f_i))$ (5)

With:

- |F|: Number of user's friends.
- Wp_i (E): weight of event E in the search history of friend f_i in the last week (calculated according to formula 1)
- W (f_i): weight of friend f_i for the user (formula 4).

3) Weighting of terms linked to a current event

The degree of global link of a term t to a current event E, denoted LGC (t, E), is given by:

$LGC(t,E) = (1/|F|)*\Sigma_{i=1..|F|}(LGP_i(t,E)*w(f_i))$ (6)

With:

- LGP_i (t, E): global link of the term *t* to the current event E in the restricted history of friend f_i (formula 3).

V. SEARCH PROCESS

The objective of our personalized search process is to improve the two important steps: (i) user query expansion. (ii) Adapting the search results to the user's interests.

Our personalized search process follows this scenario:

- Receive the user query Q, and retrieves its location.
- Select the profile base (*B_LOC, B_GNL, B_Event_P or B_Event_C*) most relevant to the query.
- The query Q is reformulated (enriched), according to the selected base.
- The reformulated query Q' is sent to one classic search engine.
- The returned results, will be subjected to another matching operation to reclassify them according to their similarity to Q'.
- The new list of reordered documents is transmitted to the user.
- The system observes the reactions and behaviors of the user to update his profile.

In the following, we will only explain treatments about *B_Event_P* and *B_Event_C*. For more detail about other two bases *B_LOC* and *B_GNL* refer to [2].

A. Selection of the profile base most appropriate to user query

To select from two profile bases *B_Event_P* and *B_Event_C*, 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:

$Sim(B\_Event\_P/Q) = \sum_{j=1..|B\_Event\_P|}\sum_{i=1..|Q|} LGP(t_i, E_j)$ (7)

$Sim(B\_Event\_C/Q) = \sum_{j=1..|B\_Event\_C|}\sum_{i=1..|Q|} LGC(t_i, E_j)$ (8)

The most similar base to user query will be selected.

B. Query reformulation

The initial user query Q will be enriched by the injection of some relevant terms and / or tags from the selected profile base. We get the reformulated query *Q'* as follows:

- Case of *B_Event_P* selected:

$Q' = Q + Ep + G$ (9)

With:

- Q = ($t_1, t_2, ... t_n$): the initial user query.
- Ep = {E_i € *B_Event_P*, such as: there is t_q € Q with LGP (t_q, E_i)> 0}: set of past events to which is linked at least one term of the initial user query.
- G = {g_j € *B_Event_P*, such as : there is E_i € Ep with LGP (g_j, E_i)> = 0) and (there is t_q € Q with *coexist(tq,gj) = 1)* } set of tags having a coexistence relation (see [2]) with at least one term of Q, and are linked at least to one element of Ep.

- Case of *B_Event_C* selected:

$Q' = Q + Ec + G'$ (10)

With:

- Ec = {E_i € *B_Event_C*, such as: there is tq € Q with LGC (tq, Ei)> 0}: set of current events to which is linked at least one term of the query.
- G' = {g_j € *B_Event_C*, such as : there is E_i € Ec with LGC (g_j, E_i)> = 0) and (there is t_q € Q with *coexist (tq, gj) = 1*) } set of tags having a relation of coexistence with at least one term of Q, and which are linked at least to an element of Ec.

C. Documents ranking

In this step, we only consider the list of N top documents returned by the classic search engine. Documents are classified according to their similarities to reformulated query *Q'* as follows

- Case of *B_Event_P* selected:

$Sim(d,Q') = \sum_{t \in Q' \cap B\_Event\_P}[(tf_d/|d|)*\sum_{E \in Ep} LGP(t,E)] * \log(N/rank(d))$ (11)

With:

- t: an element of Q '(term or tag) classified in *B_Event_P*.
- tf_d: frequency of the element *t* in the document *d*.
- |d|: size of document d.
- E: event of the Ep.
- LGP (t, E): global link of element t to event E (formula 3).
- N: number of documents in the list of tops returned documents.
- Rank (d): the order of document d in the selected list.

- Case of *B_Event_C* selected:

$Sim(d,Q') = \sum_{t \in Q' \cap B\_Event\_C}[(tf_d/|d|)*\sum_{E \in Ec} LGC(t,E)] * \log(N/rank(d))$ (12)

With:

- E: event of the Ec.

- LGC (t, E): global link of element t to event E (formula 6).

D. 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.

VI. SIMULATION

In the absence of an evaluation framework to the mobile context that represents an important part of our profile, we will evaluate our approach with a simulation in which we have differently constructed geo-social profiles for 20 users from different professional fields. Each user sends 5 queries from different locations. In sum 80 queries and 10 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 5, 10 and 15 first documents returned by Google for each of 3 following scenarios: (A) query sent without reformulation, (B) query reformulated only by geo-social user profile proposed in [2], (C) query reformulated using our proposed event-based geo-social user 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)
Pr @ 5	0.26	0.45	0.55	19 %	10 %
Pr @ 10	0.33	0.41	0.50	17 %	9 %
Pr@ 15	0.35	0.39	0.52	17 %	13 %
Pr_avg	0.313	0.416	0.523	17.66 %	10.66 %

We note that use of event-based geo-social profile has improved search results relative to the geo-social profile and the empty profile. The simulation details show that more query is short, and more the profile is rich, more our results are improved.

VII. CONCLUSION AND PERSPECTIVES

In this paper, as continuity of our works on the personalized information retrieval, we have enriched the geo-social user profile proposed in a previous work. This enrichment aims to introduce in user profile the event aspect of his interests, in order to take charge of his information needs incited by important events that occupy public opinion on the social network. The event-based extension of the geo-social profile includes: (i) the addition of two other bases to save past and current events, and their relevant related terms and tags. (ii) The establishment of another set of formulas for calculation of different weights and similarities needed for the personalized search process.

The results of our simulation show that a rich user profile can well help to improve the relevance of the search results.

As perspectives for this work, we imagine that a semantic vision to the components of the user profile (queries, terms, tags, location, events ...), and a study of the stability and the evolution modes of user profile, can bring more improvement to search results.

VIII. REFERENCES

[1] David Vallet, Miriam Fernández, Pablo Castells, Phivos Mylonas, and Yannis Avrithis , "Personalized Information Retrieval in Context using ontological knowledge", Knowledge Eng. Review 23(1): 73-100, 2008.

[2] Rafa tahar and Kechid samir, "A geo-social user profile for a personalized information retrieval", in ICIME 2016, November 02-05, 2016, Istanbul, Turkey, published by: ACM. ISBN 978-1-4503-4761-7/16/11, DOI: http://dx.doi.org/10.1145/3012258.3012270

[3] Chahrazed Bouhini, Mathias Gery, Christine Largeron. "Integrating user's profile in the query model for Social Information Retrieval". IEEE Conference on Research Challenges in Information Science, Marrakesh, Morocco 2014.

[4] Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Lucia Siciliani, "Extending an Information Retrieval System through Time Event Extraction", in 8th International Workshop on Information Filtering and Retrieval, DART 2014, Co-located with XIII AI*IA Symposium on Artificial Intelligence, AI*IA 2014 36-47 CEUR Workshop Proceedings.

[5] Ourdia Bouidghaghen., « 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.

[6] Vandenbussche, P.Y., Teissèdre, C., "Events Retrieval Using Enhanced SemanticWeb Knowledge". In Workshop DeRIVE 2011, in conjunction with 10th International Semantic Web Conference 2011 (ISWC 2011).

[7] Kanhabua, N., Nørvag, K. "Learning to Rank Search Results for Time sensitive Queries". In Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 2463–2466. CIKM '12, 2012.

[8] Arikan, I., Bedathur, S.J., Berberich, K.: "TimeWill Tell: (2009), "Leveraging Temporal Expressions in IR". In Proceedings of the 2ND International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, ACM (2009).

[9] Beatrice Arnulphy. , « Désignations nominales des événements : étude et extraction automatique dans les textes ». Phd Thesis. Univ Paris Sud - Paris XI, 2012.

[10] Ludovic Jean-Louis, Romaric Besançon et Olivier Ferret., « Une méthode d'extraction d'information fondée sur les graphes pour le remplissage de formulaires », in TAL. Volume 54 – n° 1/2013, pages 139 à 170, 2013.

[11] Laurie Serrano, Thierry Charnois Stephan Brunessaux, Bruno Grilheres et Maroua Bouzid1. (2012). « Combinaison d'approches pour l'extraction automatique d'évènements ", in Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 423–430, Grenoble, 2012.

[12] Pustejovsky J., Castaño J., Ingria R., Saurí R., Gaizauskas R., Setzer A., Katz G., « TimeML : Robust Specification of Event and Temporal Expressions in Text », IWCS-5, Fifth International Workshop on Computational Semantics., Tilburg University, 2003.

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

[14] Lamjed Ben Jabeur, Lynda Tamine et Mohand Boughanem, « Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter », in conference MARAMI 2011, Grenoble, 2011.