STRUCTURAL SIMILARITY MEASURE OF USERS PROFILES BASED ON A WEIGHTED BIPARTITE GRAPHS

I. Elachkar [1], H. Ouzif [1], H. Labriji[1]

[1] Laboratory of Technological Information and Modeling, Faculty of Sciences Ben M'sick, University Hassan II, Casablanca, Morocco - (elachkar.ibtissam, ouzif.hind)@gmail.com,labriji@yahoo.fr

KEY WORDS: User Profile, User Interests, Similarities Measures, Graph, Information Retrieval System, Recommendation System

ABSTRACT:

The user profile is a very important tool in several fields such as recommendation systems, customization systems etc., it is used to narrow the number of data or results provided for a specific user, also to minimize the cost and the time of processing of multiple systems. Whatever the user profile model used, it's updating and enrichment is a very essential step in the information research process in order to obtain more interesting and satisfactory results, which lead the information systems to develop several techniques aiming to enrich them based especially on similarity methods between user profiles. The similarity methods are used for several tasks such as the detection of duplicate profiles in online social network, also to answer the problem of cold start, and to predict users who can become friends as well as their future intentions, etc. In this paper, we propose a new approach to express the similarity between users profiles by developing a structural similarity measure to calculate the similarity between user profiles based on SimRank measure or similarity ,and the properties of bipartite graphs, in order to take advantage of the information provided by the relational structure between user profiles and their interests, our method is characterized by the similarity propagation between graph's nodes over iterations from source nodes to their successors, so our method finds profiles similar to the query profile, whether the links are direct or indirect between profiles.

1. INTRODUCTION

The fundamental purpose of information systems (IS) is to provide more satisfactory results to the needs of a given user from his query by using similarity measures to study the resemblance between this query and a collection of documents. To facilitate the processing of this task, these systems have starting to add additional information from the user to his query, such as his browsing history, his profiles on social networks, the information entered in forms, etc. A study made by (Fijałkowski, 2011) shown that the best additional information which can be integrated during information retrieval processes, is the use of the user profile, which has given rise to personalized information search system and then the contextual information search system based on user profile which integrates it in the information retrieval process such as in relevance reinjection ,query reformulation ,search results ordering ,etc. Sometimes these systems are faced profiles which do not contain all the information which can be useful for them, especially in the case of cold start problem(Lika, 2014), consequently the enrichment of these profiles is essential, the most used techniques in these cases is the processing and analysis of information of users similar to the user which we aim to complete and enrich his profile. So in this paper, we propose a new structural similarity measure, based on a weighted bipartite graph to study the similarity between profiles, since we think that the information provided by the relational structure present an interest and deserves to be studied. So in this article we will first present the user profile, its uses and some similarity measures in order to introduce our approach of structural similarity between user profiles with an application and we will end by a conclusion and our prospects for research.

1.1 The User Profile

According to (Hasan, 2013) a user profile represents a collection of personal data associated with a specific user which describes a set of attributes, these attributes may include geographic location, academic and professional experiences, objectives (short term and long term), behaviours, interests (professionals, entertainment, commercial products, etc.), etc. The user profile can be built according to two methods: either by the user himself, what is called explicit profile, or automatically from data resulting from the interactions between the user and the system, in this case it's called implicit profile. This last step is the most common, since the manual entry of parameters (preferences, interests ...) by the user can be a tiring task for him and can take a long time to express his needs.

1.2 The Use of User Profile

Users profiles are used in several areas to speed up and facilitate data processing, especially in the areas of recommendation systems such as (Alshammari, 2019) which deals with personalized recommendations on Twitter based on the explicit modeling of users profiles, as well than in the field of personalization such as the case of (Tahar, 2017) which begins a very interesting approach to information search based on semantics using a geo-social user profile, or to detect false information by exploiting the profiles of users on social networks(Shu, 2019) and extracting the opinions and interests of these users (Chen, 2017), etc.

The lifecycle of a user profile goes through several techniques, starting with the extraction of information and data of the user, then its modeling (El Achkar, 2019), its construction and finally its enrichment.

User's data changes from one moment to another, which implies a regular update of these profiles, some systems tend to exploit

the data of users similar to such a user in order to enrich his profile, which has pushed researchers to develop techniques and measures of similarity between user profiles, especially to overcome the famous problem of cold start (Lika, 2014).In the next part, we will cite some existing similarity measures, in order to introduce our similarity approach between user profiles based on a weighted bipartite graph.

1.3 Similarity Measures

There are several similarity measures in the field of information system that we can group them into 5 main types: Semantic similarity (Hliaoutakis, 2006), Structural similarity (Buttler, 2004), Content similarity (Stentiford, 2003), keyword similarity (Niwattanakul, 2013) and Hybrid similarity (Gupta, 2014), each type of similarity is exploited in a given context according to the needs and intentions of each system. For example to compare user profiles in order to enrich them, or to detect fake profiles or else for matching user profiles, also in recommendation systems in order to predict user behaviours and intentions and so on.

The most used similarity measures are: Cosine similarity (Li, 2013), Jaccard similarity (Niwattanakul, 2013),Pearson correlation coefficient (Benesty, 2009), SimRank similarity (Jeh, 2002), Aggregated similarity (Amer, 2018).

Comparative studies between these measurements show that SimRank and the Cosine measurement give satisfactory results especially in the field of collaborative filtering and another comparative study between SimRank and cosine conducted by (Champclaux, 2008) in the field of information retrieval, demonstrates that the SimRank outperform, which motivates our approach to apply the SimRank measure on user profiles in order to study the similarity between them using a weighted bipartite graph.

2. OUR APPROACH

Our work revolves around the development of a structural similarity measure to calculate the similarity between users profiles, this similarity is based on the structural measure of similarity SimRank (Jeh, 2002). This part is organized as follows: we will start with the presentation of the SimRank similarity measurement based on an oriented graph, as well as the generic bipartite SimRank measurement in order to introduce our approach and the methodology that we will follow to measure the similarity between two user profiles.

2.1 The SimRank Model Based on an Oriented Graph

(Jeh, 2002) Proposed a measure of structural similarity between objects in a domain involving an object-to-object relationship. In this approach, the objects and their relations are modeled by an oriented graph G(V, E), which the nodes V represent the domain objects studied, and the arcs E represent the relations between these objects.

The initial assumption is that "objects are similar if they are connected by similar objects". The aim of this approach is to determine the similarities between nodes of the graph by assigning them a similarity score called SimRank which is defined by:

Let I(v) the set of predecessors of a node v, |I(v)| is the cardinal of all these predecessors. The SimRank Score S(a, b) between an object a and an object b is defined by:

$$If\ a = b : S(a,b) = 1 \qquad (1)$$

$$Else : \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S(I_i(a), I_j(b))$$

2.2 The SimRank Model Based on Bipartite Graph

The SimRank measure was extended by (Jeh, 2002) to fields with two types of objects. The appropriate structure to represent such a domain is a bipartite graph. So we can calculate two types of similarity scores:

_ The similarity score between nodes of type 1: Two object of type 1 are considered similar if they point to similar objects of type 2.

_ The similarity score between type 2 objects, two objects of type 2 are similar if they are pointed by similar type 1 objects.
These notions can be formalized by two functions S_1 and S_2, the SimRank Score S(a, b) between two objects a and b is defined by:

$$S_1(a,b) = \frac{C_1}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} S_2(O_i(a), O_j(b)) \qquad (2)$$

$$S_2(a,b) = \frac{C_2}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S_1(I_i(a), I_j(b))$$

Where:

O(a) : is the set of successors of node a, and I (a) is the set of its predecessors.
|O(a)|: is the cardinal of the set of successors and |I(v)| is the cardinal of all the predecessors.
C_1 and C_2 are constants between 0 and 1

Experiments of these formulas by (Champclaux, 2009) have shown that this measure of similarity is characterized by the capacity to order objects according to their relationships as well as the illustration of similarity propagation phenomenon which is the basis of our approach for studying similarity between user profiles from their interests.

2.3 Our Structural Similarity Approach between User Profiles Based On Weighted Bipartite Graphs

As we mentioned earlier, the user profile is considered as a set of data made up of a various information: personal, professional, and especially user's interests which we will use in our study. The application of the structural measurement of SimRank described above consists on one hand in representing this data in the form of a bipartite graph in which the type 1 nodes are user profiles and the type 2 nodes are the interests of these users, and secondly to define the structural relationship between them: The belonging, that is to say the fact that a profile contains interests and vice versa that the interests are contained in a profile, a profile node is connected by an arc to an interest node if the profile contains this interest, and finally searching user profiles similar to a given user profile by the application of Simrank. A query that contains the user profile in

which we are searching for profiles those simulate it, is integrated in this graph as an additional profile node.

Example: Considering a corpus of research composed of two profiles made up of a set of interests like that:

_ Profile1: {interest1, interest2, interest4, interest5}
_ profile2: {interest2, interest3, interest5}

Given a search query: R-Profile: {interest1, interest3, interest5} that represents the user profile which we are searching profiles those similar to him.

The corpus and the query are represented by the following graph G:

G({Profil1,Profil2,R-Profile};={interest1,interest2,interest3, interest4,interest5};E={(Profile1,interest1);(Profile1,interest2);(Profile1,interest4);(Profile1,interest5);(Profile2,interest2);(Profi le2,interest3);(Profile2,interest5) ;(R-Profile,interest1);(R-Profile,interest3);(R-Profile, interest 5)})

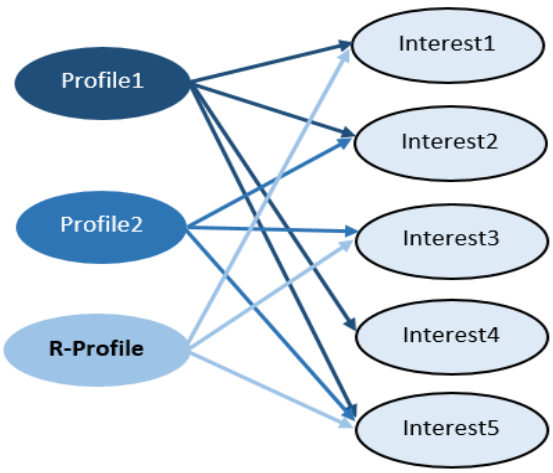

Figure 1. Profiles-Interests Bipartite Graph

Our goal is to sort profiles based on their similarity to the R-profile query.

2.4 The Formula

In Information retrieval field, the best results are obtained when documents are represented in the form of a weighted terms list that is why we want to adopt this principle and add the weight of users interests in this approach. Such a description is translated by a weighted bipartite graph in which the arcs between profiles nodes and interests' nodes are weighted by the weight of these interests appearing in each profile. So the SimRank formulas adapted to our approach will be presented as follows:

Considering a corpus described by: C and P where:
C=(cj), j=1..m : is the set of corpus's interests, m is the total number of these interests.

P=(pi), i=1..n : is the set of corpus's profiles, n is the total number of profiles in this collection.

With pi=(wi1,wi2,…,wij,…wim) , wij is the weight of the interest j in the profile i. In order to take into account interest's weights, in the intention of giving the Profiles-interest arcs a weight.

The calculation of the similarity $S_p(p_i, p_j)$ between two profiles $p_i \; and \; p_j$ is defined as follows:

$$If\ i = j : S_p(P_i, P_j) = 1 \qquad (3)$$

$$Else: S_p(P_i, P_j) = \begin{cases} \frac{M}{|C_{P_i}||C_{P_j}|} \sum_{c_k \in C_{p_i}} \sum_{c_l \in C_{p_j}} w_{ik} \times w_{ik} \times S_C(c_{k(p_i)}, (c_{l(p_j)}) \end{cases}$$

Where :

M : a propagation constant M= 0,9

C_{P_i} : Is the set of interests of the profile Pi.

$|C_{P_i}|$: Is the number of interests belonging to the profile Pi.

$c_k(p_i)$: Is the k^{th} interest of the profile Pi (the i^{th} profile of the collection).

The similarity $S_c(c_i, c_j)$ between two interests c_i and c_j is defined as follows:

$$If\ i = j : S_c(c_i, c_j) = 1 \qquad (4)$$

$$Else: S_c(c_i, c_j) = \begin{cases} \frac{M}{|P_{c_i}||P_{c_j}|} \sum_{p_k \in p_{c_i}} \sum_{p_l \in p_{c_j}} w_{ki} \times w_{lj} \times S_p(p_{k(c_i)}, (p_{l(c_j)}) \end{cases}$$

Where:

Pc_i: Is the set of profiles containing the interest c_i.

$|Pc_i|$: Is the number of profiles containing the interest c_i.

$p_k(c_i)$: Is the k^{th} profile containing the interest c_i.

The formulas reflect the fact that the similarity of two profiles strongly depends on the similarity of the interests that contain them and reciprocally the similarity of two interests depends on the similarity between the profiles in which they belong, this is due to the structural relationship between each profile and its interests

3. APPLICATION

In this part, we will apply the formulas presented previously to a corpus composed of three profiles (P1, P2, and P3) and a query (R-Profile). The R-Profile query is composed of five interests: int1, int2, int3, int4, int5. The P1 profile is composed of five interests: int1, int2, int3 that it shares with the query (R-Profile), plus int6 and int7. The P2 profile is composed of four interests: int4 and int5 which it shares with the request (R-Profile), plus int8 and int9, the P3 profile is composed of three interests: int6 and int7 which it shares with the P1 profile, and int8 which it shares with the P2 profile, and finally the P4 profile which contains the interests int10 and int11. int1, int2 and int3 have a weight of 2, the other interests have a weight of 1. This example is illustrated by the following figures:

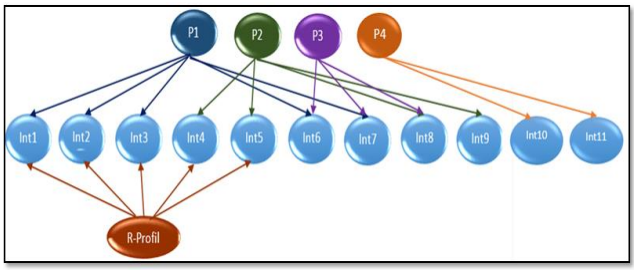

Figure 2. Graph Representing the Profiles P (1, 2, and 3) and the Query R-Profile

	Int1	Int2	Int3	Int4	Int5	Int6	Int7	Int8	Int9	Int10	Int11
P1	1	1	1	0	0	1	1	0	0	0	0
P2	0	0	0	1	1	0	0	1	1	0	0
P3	0	0	0	0	0	1	1	1	0	0	0
P4	0	0	0	0	0	0	0	0	0	1	1
R-profile	2	2	2	1	1	0	0	0	0	0	0

Figure 3. Profiles-Interests Matrix

Here are the similarity scores of our approach, the SimRank similarity and the Cosine similarity that each profile obtains with the R-Profile query:

Profiles	Our Approach	SimRank Similarity	Cosine Similarity
P1	0,421	0.588	0,717
P2	0,325	0.438	0,267
P3	0,247	0.446	0
P4	0	0	0

Table 1. SimRank and Cosine Scores

According to the results the profile P1 is the most relevant for the query in front of P2, itself in front of P3. If we apply the cosine measurement between the query and profile 3, we get a similarity score of 0 since there is no common term between them, contrary to our approach where we obtain a score of 0.27, this is due to the direct resemblance of P3 with P1 and P2, and to the direct resemblance of P1 and P2 with R-Profile, these transitive relations reflects the phenomenon of similarities propagation between the corpus profiles, which gives strength to our approach.

We also notice that we obtained similar scores between SimRank and our approach since our approach is an extension of the SimRank similarity to which we add the interest weights of the profiles

3.1 Analysis and Discussion

We have described the adaptation of an objects comparison method based on graphs to the comparison of users' profiles. This adaptation resulted in the definition of a new similarity function taking into account the graph structure induced by the relationship between profiles and their interests. Conceptually a profile is considered as the node of a bipartite graph to which the interest nodes are connected. The similarity between profiles is calculated as the average of interests' similarities that compose them. Reciprocally, the similarity between interests is calculated as the average of profiles similarities that contain them. From this recursive definition, we defined two formulas: one defining the similarity between the profiles, the other defining the similarity between the interests, which allowed us to define a measure of structural similarity inter-profiles and inter-interests.We also took into account the profiles interests weighting, which translates conceptually by weighting the graph arcs between the profiles and their interests.

Our similarity approach high score in comparing two profiles depends more on the proportion of common interests than on the proportion of non-common interests. In addition to this, our similarity approach finds profiles similar to the query(R-profile), whether the link is direct or indirect between them, since this algorithm propagates similarities between profiles over iterations from node to node, from source nodes to their successors.

4. CONCLUSION AND PERSPECTIVES

We have presented a structural similarity measure between users' profiles, able to extract similar profiles even if there is no link or common interest between them. our approach can be used by several domains, for example it can be used to solve the famous cold start problem (Lika, 2014), also to study community evolution graph. In our case, we aim to take advantage of the similarity propagation property of our approach in order to detect nodes that can be connected in the future in a given network, for example in the case of friends networks to predict the users who can become friends as well as their future intentions, especially since it is quite obvious that a user will certainly be influenced by the interests and behaviors of the users of his network.

REFERENCES

Alshammari, Abdullah, Nikolaos Polatidis, Stelios Kapetanakis, et Roger Evans. 2019. « Personalized Recommendations on Twitter Based on Explicit User Relationship Modelling »

Amer, Ali A., Marghny H. Mohamed, Adel A. Sewisy, et Khaled Al Asri. 2018. « An Aggregated Similarity Based Hierarchical Clustering Technique for Relational DDBS Design ». In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), 295‑99. https://doi.org/10.1109/PDGC.2018.8745981.

Benesty, Jacob, Jingdong Chen, Yiteng Huang, et Israel Cohen. 2009. « Pearson Correlation Coefficient ». In Noise Reduction in Speech Processing, édité par Israel Cohen, Yiteng Huang, Jingdong Chen, et Jacob Benesty, 1‑4. Springer Topics in Signal Processing. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-00296-0_5.

Buttler, David. 2004. « A Short Survey of Document Structure Similarity Algorithms »

Champclaux, Yaël. 2009. « Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information ». Phdthesis, Université Paul Sabatier - Toulouse

Champclaux, Yaël, Taoufiq Dkaki, et Josiane Mothe. 2008. « Enhancing High Precision Using Structural Similarities »

Chen, Hongxu, Hongzhi Yin, Xue Li, Meng Wang, Weitong Chen, et Tong Chen. 2017. « People Opinion Topic Model: Opinion based User Clustering in Social Networks ». In Proceedings of the 26th International Conference on World

Wide Web Companion, 1353–1359. WWW '17 Companion. Perth, Australia: https://doi.org/10.1145/3041021.3051159.

El Achkar, Ibtissam, Amine Labriji, et Labriji El Houssine. 2019. « A Semantic Method to Extract the User Interest Center ». In Innovations in Smart Cities Applications Edition 2 522‑34https://doi.org/10.1007/978-3-030-11196-0_44.

Fijałkowski, Damian, et Radosław Zatoka. 2011. « An architecture of a web recommender system using social network user profiles for e-commerce ». In 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), 287‑90.

Gupta, Yogesh, Ashish Saini, et AK Saxena. 2014. « Fuzzy Logic-Based Approach to Develop Hybrid Similarity Measure for Efficient Information Retrieval ». Journal of Information Science.https://doi.org/10.1177/0165551514548989.

Hasan, Omar, Benjamin Habegger, Lionel Brunie, Nadia Bennani, et Ernesto Damiani. 2013. « A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case ». In 2013 IEEE International Congress on Big Data, 25‑30. https://doi.org/10.1109/

Hliaoutakis, Angelos, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis, et Evangelos Milios. 2006. « Information Retrieval by Semantic Similarity » International Journal on Semantic Web and Information Systems (IJSWIS).

Jeh, Glen, et Jennifer Widom. 2002. « SimRank: a measure of structural-context similarity ». In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 538–543. https://doi.org/10.1145/775047.775126.

Li, Baoli, et Liping Han. 2013. « Distance Weighted Cosine Similarity Measure for Text Classification ». In Intelligent Data Engineering and Automated Learning – IDEAL 2013 https://doi.org/10.1007/978-3-642-41278-3_74.

Lika, Blerina, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. « Facing the Cold Start Problem in Recommender Systems ». Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2013.09.005.

Niwattanakul, Suphakit, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. « Using of Jaccard Coefficient for Keywords Similarity ».

Shu, Kai, Xinyi Zhou, Suhang Wang, Reza Zafarani, et Huan Liu. 2019. « The role of user profiles for fake news detection ». In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 436–439. ASONAM '19. Vancouver, British Columbia, Canada: https://doi.org/10.1145/3341161.3342927.

Stentiford, Fred W. M. 2003. « Attention-based similarity measure with application to content-based information retrieval ». In Storage and Retrieval for Media Databases 2003, https://doi.org/10.1117/12.476255.

Tahar, Rafa, et Kechid Samir. 2017. « An Event-Based Geo-Social User Profile for a Personalized Information Retrieval ». First International Conference on Embedded & Distributed Systems: IEEE.https://doi.org/10.1109/EDIS.2017.8284030.