FUSION METHODS TO FIND WEB COMMUNITIES

João Ferreira
Instituto Superior de Engenharia de Lisboa
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, Instituto Superior Técnico
alberto.silva@acm.org

José Delgado
Instituto Superior Técnico
jose.delgado@tagus.ist.utl.pt

ABSTRACT

We propose a new method of community identification based on the combination of different approaches (context and link analyses) using stable information user's needs (profiles) and documents (as author's profile). We discuss application, standards and identify gaps and needs in the process of web community's identification.

KEYWORDS

Web Community, fusion, vectorial, link analyses, clustering and community.

1. INTRODUCTION

Information on the web is available to millions of different individuals, operating independently, and having a variety of backgrounds, knowledge, goals, and cultures. Web has a decentralized, unorganized, and heterogeneous nature. Communities are a way of, pre-defined groups of persons with similar interests, organize and diffuse information. There are 3 major questions facing designers of on-line communities: how to get users to behave, how to get users to contribute with content of quality, and how to get users to return and contribute on an ongoing basis. This paper discusses fusion of techniques (context and link analyses) to identified Web communities in order to use the best of each technique and discuss problem and the needs of standards.

2. WEB COMMUNITIES

Web Communities is defined as a group of users that share common interests and can be of two types: (1) of users with information needs (established from user profiles or log file of user activity in a search engines); (2) of authors taken from de documents available (document cluster), which can be representative of authors profile. To accomplish this concept we propose a web community build from the fusion of 3 approaches:

- (1) Stable user's information needs; clustering user profiles and central profile identified as community profile.
- (2) from authors (document clusters), central cluster document is identified and used as community profile to combine with (1).
- (3) from links of community relevant documents and user browsing. This link information provides more information (terms \ concepts) to final community profile.

This different approach enriches Web communities profiles based on the different vision of authors and users, using textual and link information.

2.1 Cluster analysis

The first step on community's identification is the recognition of a cluster of profiles and documents on a Vector Space Model (VSM) using the SMART length-normalized term weights as implemented through WebSearchTester [1]. From text, tags are removed; stop words and weights are based on *Lnu* document term weight (1):

$$d_{ik} = \frac{(1+\log(f_{ik}))/(1+\log(avg\_f_i))}{(1.0-slope)*p_i + slope*t} \quad (1); \quad q_k = \frac{(\log(f_k)+1)*idf_k}{\sqrt{\sum_{j=1}^{t}\left[(\log(f_j)+1)*idf_j\right]^2}} \quad (2);$$

$$\mathbf{q}^{\mathrm{T}} q_i = \sum_{k=1}^{t} q_k q_{ik} \quad (3); \quad d^{\mathrm{T}} q_i = \sum_{k=1}^{t} d_k q_{ik} \quad (4)$$

with the slope of 0.3 for document terms [2], where f_{ik} is the number of times term k appears in document i; avg_f_i is the average in-document frequency for document i; t is the number of unique terms in the collection and p_i is the average number of unique terms in a document i. The formula for *ltc* profile term weight is (2), where f_k is the number of times term k appears in the profile; and idf_k is the inverse document frequency [3] of term k. The denominator is a document length normalization factor, which compensates for the length variation in queries. Documents and profiles were clustered by their inner product (4) and (3).

2.2 Fusion

The Similarity Merge (SM) fusion formula, originally introduced by Fox and Shaw [4,5] and refined by Lee [6,7], computes the fusion score of a document by the sum of normalized component scores boosted by the retrieval overlap. SM fusion formula is used to merge and rank user and authors profiles: $F = \sum_{i=1}^{2} NS_i * olp$, (5); where F = fusion score, NS_i = normalized score of a document by system i, olp = 1 if profile belongs to approach one (user's) or two (author's), or 2 if profile belongs to both. The normalized document score, NS_i, is computed by Lee's min-max formula [7,8], where S_i is the retrieval score of a given document and S_{max} and S_{min} are the maximum and minimum profiles scores by system i: $NS_i = (S_i - S_{min}) / (S_{max} - S_{min})$, (6)

2.3 Link analyses

From a set of relevant documents of each community profile, link analyses were performed on this set of documents to find other similar and then find new terms.

We implemented [8], which essentially normalize the contribution of authorship by dividing the contribution of each page by the number of pages created by the same author, was used to modify the HITS formulas as follows:

$$a(p) = \sum_{q \to p} h(q) \times auth\_wt(q,p), \quad (7) \qquad h(p) = \sum_{p \to q} a(q) \times hub\_wt(p,q), \quad (8)$$

In above equations, *auth_wt(q,p)* is 1/m for page q, whose host has m documents pointing to p, and *hub_wt(p,q)* is 1/n for page q, which is pointed by n documents from the host of p. From this process a new set of relevant documents were identified and from Top ten positive and top two negative weighted terms from the top three ranked documents of the initial retrieval results were used to expand the fusion community profile in a pseudo-feedback retrieval process based on the adaptive linear model. The basic approach of the adaptive linear model, which is based on the concept of preference relations from decision theory [9], is to find a *solution vector* that will rank a more-preferred document before a less-preferred one [10].

The solution vector is arrived at via an *error-correction procedure*, which begins with a *starting vector* $\mathbf{q}_{(0)}$ and repeats the cycle of "error-correction" until a vector is found that ranks documents \ profiles according to the preference order estimation based on relevance feedback [11]. The error-correction cycle i is defined by $\mathbf{q}_{(i+1)} = \mathbf{q}_{(i)} + \alpha\mathbf{b}$ (9) where α is a constant, and $\mathbf{b}$ is the *difference vector* resulting from subtracting a less-preferred document vector from a more preferred one [12]. The choices for the constant α and the *starting vector* $\mathbf{q}_{(0)}$ are ad-hoc.

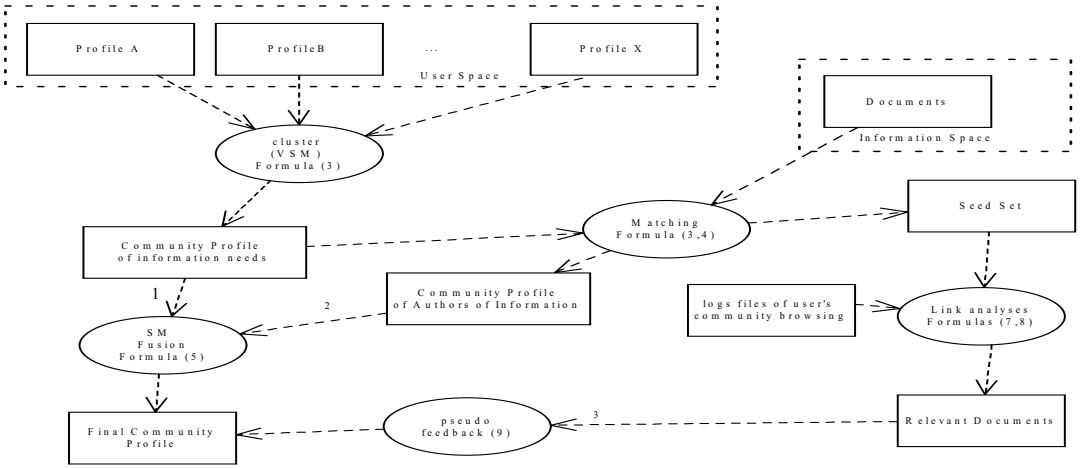

Figure 1. Process of community identification.

3. APPLICATION AND STANDARDS COMMUNITIES

From the process described on previous section results of each community identified a list of terms with weights. Next step is the identification of central profile which is used as community representative and will be chosen as central profile the nearest to the geometric centre of the community. The communities of users will be useful information to different organisations (publishers, editors, etc). The central profile of a given community can be used to define "special" media services, like personalised newspaper, digital TV and others.

The concept of community can be used as a free approach that later can influence the dynamic creation of information catalogues.

Complex communities can be divided into smaller groups inside a small sub-space. Community's boundaries will be an interesting and difficult problem to solve, where users can belong to one or more communities (see Figure 2). All this measures are performed in an N dimensional Euclidean space where profiles are defined by vectors. The dimension of space is defined by classified terms available in document collections. We applied this concept of web communities into two study cases: MyTv [13], personalized filter system to identify Tv programs and MyEnterpriseNews [13] a filtering system to identify enterprise news related over the Web.

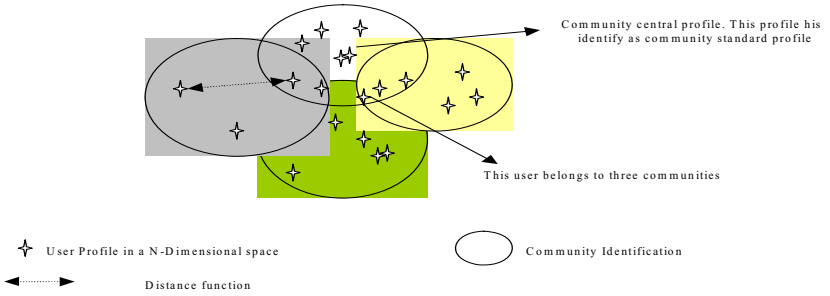

Figure 2. Web Community central profile.

4. CONCLUSION AND FUTURE WORK

Establish Web Communities is more than find best techniques or algorithms, but also a problem that needs consensus (how to do and how to validate communities), standards and acceptance from the big web community. Similar of the efforts made on standardization performed (like Web Semantics, Soap, OIL) on the documents side, we will need equivalent on the user's side. It is a complementary work that needs to be performed, because Web is not documents but also users.

Other important problem is the validation of web communities. There are no controlled environments to test systems and performed measures and then compare performances of different systems.

Interesting will be once these web communities are establish and defined check and analyze differences between them and the knowledge space previous build (e.g. classification systems). On the document side we can compare web directories (web community of document) against classified systems.

There is a need of standards automatic process to identify web community and put that information available. Growth of Internet can be influence from these communities, if they are known.

It is also possible to combine classification systems in the process of community identification, to merge in the process human knowledge previous build.

REFERENCES

[1] Ferreira, João; Silva, Alberto; Delgado, José (2004). Infraestrutura modular de teste para pesquisa de informação. *Proceedings of Ibero-Americana WWW/Internet 2004*, Madrid – 7th to 8th October 2004.

[2] Buckley C. Singhal A. e Mitra M. (1997). Using query zoning and correlation within SMART: TREC 5. In E. M. Voorhees e D. K. Harman (Eds.) The Fifth Text REtrieval Conference (TREC-5) (NIST Spec. Publ. 500-238 pp. 105-118). Washington DC: U.S. Government Printing Office.

[3] Sparck J. K. (1971). Automatic Keyword Classification for Information Retrieval. London: Butterworth.

[4] Fox E. A. & Shaw J. A. Combination of multiple searches. In D. K. Harman (Ed.). TREC-2, 1994.

[5] Fox, E. A., & Shaw, J. A. Combination of multiple searches. In D. K. Harman (Ed.), *The Third Text Rerieval Conference (TREC-3)* Washington, DC: U.S. Government Printing Office, 1995, NIST Spec. Publ. 500-225, 105-108.

[6] Lee J. H.. *Combining multiple evidence from different relevance feedback methods (Tech. Rep. No. IR-87).* Amherst: University of Massachusetts Center for Intelligent Information Retrieval, 1996.

[7] Lee J. H. Analyses of multiple evidence combination. *Proceedings of the ACM SIGIR Conference on Research and Development in IR*, 1997, 267-276.

[8] Bharat K. e Henzinger M. R. (1998). Improved Algorithms for Topic Distillation in Hyperlinked Environments. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval 104-111.

[9] Fishburn, P. C. (1970). *Utility theory for decision making*. New York: John Wiley & Sons.

[10] Wong, S. K. M., Yao, Y. Y., & Bollmann, P. (1988). Linear structure in information retrieval. *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 219-232.

[11] Wong, S. K. M., Yao, Y. Y., Salton, G., & Buckley, C. (1991). Evaluation of an adaptive linear model. *Journal of the American Society for Information Science*, *42*, 723-730.

[12] Sumner, R. G., Jr., Yang, K., Akers, R., & Shaw, W. M., Jr. (1998). Interactive retrieval using IRIS: TREC-6 experiments. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6)*.

[13] Ferreira, João; Silva, Alberto; Delgado, José (2005). Personalized Filtering Systems based on the Combination of different Methods. *Proceedings of Applied Computing 2005, IADIS, Algarve 22th to 25th February 2005.*