Profiling Users to Perform Contextual Advertising

Andrea Addis
Dept. of Electrical and Electronic Engineering
University of Cagliari
Italy
addis@diee.unica.it

Giuliano Armano
Dept. of Electrical and Electronic Engineering
University of Cagliari
Italy
armano@diee.unica.it

Eloisa Vargiu
Dept. of Electrical and Electronic Engineering
University of Cagliari
Italy
vargiu@diee.unica.it

ABSTRACT

Recommendation systems are typically aimed at proposing items to users. Nevertheless, another way of performing recommendations is viable, i.e. proposing users to domain specific web sites. Within this context, users require to be represented according to their preferences –given in terms of categories of interest. To better highlight the need for "recommending users", let us recall that commercial web sites are typically involved with one or more domain-specific businesses. In this scenario, a system able to identify relevant categories can be useful to identify users that may become target for advertisement (for instance, a company that sells pet supplies, food and products is interested in identifying users that have the "Animals" category among their interests). The goal of this research is to develop agent-based referral systems for user profiling able to identify user interests. Given a taxonomy and a set of documents representing a user (i.e. selected by the user while surfing the web) the system is able to profile her/him in terms of the given categories. For the sake of simplicity, experiments have been performed using WordNet Domains as reference taxonomy and Wikipedia as document source. The underlying assumption is that, due to the general-purpose machine learning techniques adopted to implement the system, this capability is exportable to other –more specific– taxonomies.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, selection process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*user profiles and alert services*

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Most of the advertisements on the web are in fact short textual messages usually marked as "sponsored links". Two main kinds of textual advertisements can be highlighted on the web today [4]: (i) sponsored search advertising; and (ii) contextual advertising. The former puts advertisements on the pages returned from a web search engine following a query. All major current web search engines support such ads and act simultaneously as search engine and advertisement agency. The latter puts advertisements within the content of a generic, third party, web page. Usually a commercial intermediary, namely an ad-network, is in charge of optimizing the selection of advertisements with the twin goal of increasing revenue (shared between publisher and ad-network) and improving user experience. In other words, contextual advertising is a form of targeted advertising for advertisements appearing on websites or other media, such as content displayed in mobile browsers. The advertisements themselves are selected and served by automated systems based on the content displayed to the user.

In our opinion, another issue to be taken into account is how to personalize advertisements according to user preferences and interests. In this view, users can be proposed to domain specific web sites in order to provide them interesting advertisements, giving rise to suitable "user recommendation systems". Within this context, users require to be represented according to their preferences –given in terms of categories of interest.

The remainder of the paper is organized as follows: first, an overview on user profiling is given. Subsequently the proposed approach to profile users according to their interests expressed in terms of categories is described. Then, experiments results are presented. Conclusions and future directions end the paper.

2. USER PROFILING

People find hard articulating what they are looking for, but they are very good in recognizing it when they see it [9]. This insight has led to the utilization of relevance feedback, where people rate items as "interesting" or "uninteresting" and the system tries to find items that match (i.e., positive examples) and do not match (i.e., the negative examples). With sufficient positive and negative examples, modern machine learning techniques can classify new items with impressive accuracy [6].

User profiling is typically either knowledge-based or behavior-based. Knowledge-based approaches engineer static models of users and dynamically match users to the closest model. Questionnaires and interviews are often employed to obtain this user knowledge. Behavior-based approaches use the user behavior as a model, commonly using machine-learning techniques to discover useful patterns in the behavior. Behavioral logging is employed to obtain the data necessary from which to extract patterns. Kobsa [5] provides a good survey of user modelling techniques.

The user profiling approach used by most recommender systems is behavior-based, commonly using a binary class model

to represent what users find interesting and uninteresting. Machine-learning techniques are then used to find potential items of interest in respect to the binary model. There are a lot of effective machine learning algorithms based on two classes. A binary profile does not, however, lend itself to sharing examples of interest or integrating any domain knowledge that might be available. Sebastiani [13] provides a good survey of current machine learning techniques.

The choice of the most suitable algorithm for a recommender system depends on many issues, including the specific type of the service, the nature of items, as well as kind and amount of available information. For instance, if items are documents, an algorithm based on information retrieval is more appropriate because it is able to deal with problems related to the automatic analysis of text ([15], [14]). If items are multimedia with scarce descriptions, but ratings issued by the community of clients, a collaborative filtering could be more suitable ([12], [7]). When additional textual information is available (such as title, description, etc.), it is possible to combine these approaches, thus implementing a hybrid system, able to typically outperform its components ([3], [2]).

3. A CONTEXT-BASED USER PROFILER

As we are interested in classifying user preferences into selected categories belonging to a given taxonomy, we decided to adopt a content-based approach. For each user of the system, a profile is generated from a set of documents rated as relevant by the selected user, i.e., the user history[1]. New documents can then be proposed to the user and added to her/his user history if they match the computed profile. The algorithm integrates statistical and semantic approach. Figure 1 sketches the architecture of the proposed system, composed by four main modules: statistical document analyzer, semantic words analyzer, semantic network handler, and profiler.

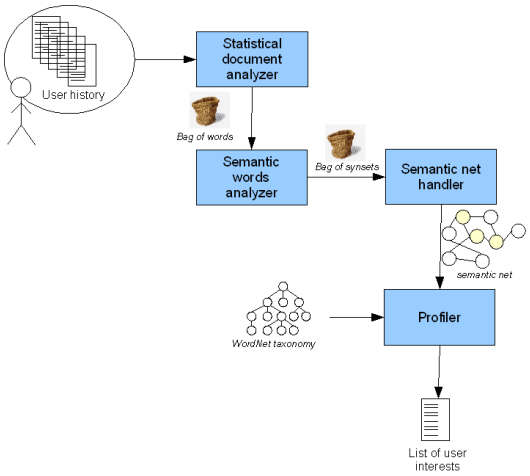

Figure 1: User profile generation at a glance.

[1]Being interested in creating user profiles, in this paper we are not interested in the way the user history is actually built.

synsetID: *02001223*	
set of synonyms: dog, domestic dog, Canis familiaris	***meaning***: a member of the genus Canis that has been domesticated by man since prehistoric times.
synsetID: *09465341*	
set of synonyms: frump, dog	***meaning***: a dull unattractive unpleasant girl or woman.
synsetID: *09382160*	
set of synonyms: dog	***meaning***: informal term for a man.
synsetID: *09256536*	
set of synonyms: cad, bounder, blackguard, dog, hound, heel	***meaning***: someone who is morally reprehensible.
synsetID: *07205647*	
set of synonyms: frank, frankfurter, hotdog, dog, wiener, wienerwurst, weenie	***meaning***: a smooth-textured sausage of minced beef or pork usually smoked.
synsetID: *03754154*	
set of synonyms: pawl, detent, click, dog	***meaning***: a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward.
synsetID: *02617005*	
set of synonyms: andiron, firedog, dog, dog-iron	***meaning***: metal supports for logs in a fireplace.

Figure 2: Synsets of the word *dog*.

Statistical document analyzer. While analyzing documents rated as relevant by the user (i.e., the user history), this module is devoted to create the bag of words (BoW), aimed at collecting all terms contained in the input texts, suitably weighted. The statistical document analyzer removes from the BoW all non-informative words such as prepositions, conjunctions, pronouns and very common verbs using a stop-word list. Subsequently, it calculates the weight of each term adopting the TFIDF measure [11]. Let us recall that, for each term t_k, the TFIDF determines the weight w_{jk} of t_k in the document d_j.

Being D the user history, the statistical document analyzer calculates an overall TFDIF considering all documents in D according to the formula:

$$tfidf_D(t_k, D) = \frac{\sum_j tfidf_D(t_k, d_j)}{\#(D, t_k)} \quad (1)$$

where $\#(D, t_k)$ denotes the number of documents in D in which the term t_k occurs at least once (also known as the document frequency of t_k), and $tfidf_D(t_k, d_j)$ is the classical TFIDF value of t_k with respect to d_j (measured in D). In symbols:

$$tfidf_D(t_k, d_j) = \#(t_k, d_j) \cdot log\frac{|D|}{\#(D, t_k)} \quad (2)$$

where $\#(t_k, d_j)$ denotes the number of occurrences of t_k in d_j and $|D|$ is the cardinality of the user history D. Furthermore, the weights resulting from TFIDF undergo a cosine normalization, given by:

$$w_{jk} = \frac{tfidf_D(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T_D|}(tfidf_D(t_s, d_j)^2}} \quad (3)$$

where T_D is the set of significant terms that appear in the set of documents D.

To reduce the dimensionality of the space, only the first

N terms of the BoW are retained. The optimal value of N must be calculated experimentally; in this work, good values have been found in the range 60-90 (due to the fact that –here– textual descriptions are often short). Hereinafter, the set of terms stored in the BoW will be called *features*.

Semantic words analyzer. This module creates the bag of synsets (BoS), which collects all synsets related to the selected features. To this end, the semantic document analyzer queries the online lexical database WordNet [10]. In particular, for each feature, WordNet provides all the corresponding synsets. As an example, Figure 2 shows all synsets of the term *dog*, each reporting a proper ID, the corresponding synonyms, and the meaning. After synset extraction, the semantic document analyzer assigns each sysnset a weight according to the TFIDF of all related terms.

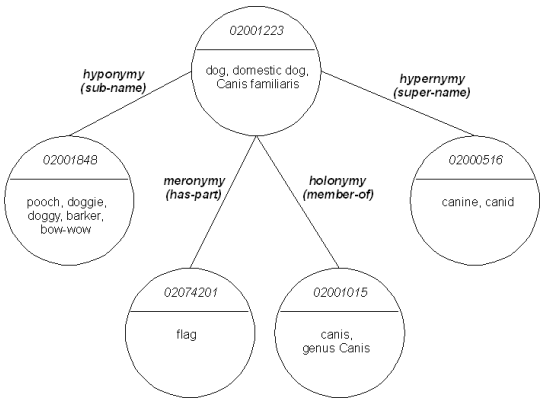

Figure 3: Semantic relations for the synsets of the word *dog*.

Semantic net handler. This module aims (i) to build the semantic net from the BoS and (ii) to extract its most relevant nodes. First, the semantic net is built in form of a graph, where nodes are the synsets belonging to the BoS, and edges are semantic relations between synsets. Four kinds of semantic relations are taken into account: *hyponymy* (sub-name) and its inverse, i.e., *hyperonymy* (super-name); *meronymy* (has-part) and its inverse, i.e., *holonymy* (member-of). Figure 3 shows an example for the synset *02001223* of the word *dog*. Then, the semantic net handler prunes the network by dropping not relevant nodes, identified according to their weight and to the number of connections with other nodes.

Profiler. This module is devoted to extract the user profile. To this end, it exploits the WordNet domain hierarchy (WNDH) [8] and associates the proper category to each selected node For example, let us note that for the synset *02001223* the corresponding domains are *Animals* and *Biology*. Considering the selected nodes, together with their weights, the profiler is able to identify the real interests of the user in terms of WordNet domains. In particular, the user profile is represented as a set of pairs $< c_k, w_k >$, where c_k are the WordNet domain categories and w_k are the corresponding weights in the range $[0, 1]$.

Choosing a suitable taxonomy that represents advertising contexts instead of WNDH, the approach can be easily applied to profile users that may become target for advertisement.

4. EXPERIMENTAL RESULTS

To assess the effectiveness of the proposed approach, experiments have been done to measure the performances in terms of mean square error (MSE).

Waiting for applying the proposed approach in the field of advertisement, experiments have been performed using WordNet Domains as reference taxonomy and Wikipedia as document source. The adoption of the latter as document source is due to the fact that Wikipedia pages have a standard structure that allows to isolate the text describing the main concepts of each topic to assign few specific categories to each page. To put into evidence that our approach can be adopted to profile users in terms of the categories they are interested in, first we resorted to the approach described in [1] to build a dataset in which documents are classified according to WordNet Domains categories. Selecting documents from such dataset allowed us to automatically create user histories.

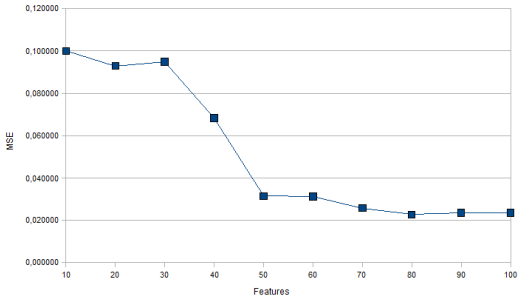

Figure 4: Performances obtained varying the number of features selected by the Statistical Document Analyzer.

Several experiments have been performed, averaging on the number n of categories (ranging from 1 to n) and keeping fixed (at a rate $1/n$) the amount of documents belonging to each category. Figure 4 reports the results obtained for different numbers N of features (from 10 to 100) and shows that the best result is obtained for $N = 80$.

Subsequently, we performed also a preliminary study about the impact of changing the number of documents associated to a given category in a user history. As a starting point for studying such phenomenon (say category imbalancement), only two categories have been considered. In particular, for each combination of two categories (say, A and B), experiments have been performed starting from 5% of documents belonging to A and 95% to B, incrementing such percentage up to 95% for A and vice versa for B (with delta=5%). Comparing such percentages with those given as output by the system, we calculated the mean square error. Figure 5 illustrates results obtained by averaging the results belonging to all combinations. The figure puts into evidence that the performance of the system depends also on category imbalancement. In particular, experimental results point out that the filtering activity of the Statistical Document Analyzer is more effective when the user history is composed by

a large amount of documents of a specific class. In our view, this is due to the fact that a strong bias on a given category facilitates the system in the task of identifying it. Moreover, the fact that in this case the remaining category has a lower range of variation has also a positively impact. Let us also note that this result is obtained by averaging all pairwise combinations of categories, without taking into account that some categories may be (and actually are) correlated (e.g. *Medicine* and *Biology*).

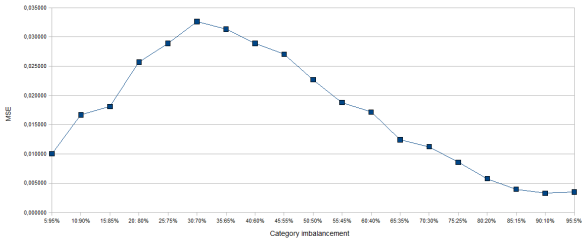

Figure 5: Performances of the system while varying the category imbalancement.

Summarizing, our preliminary experimental results confirm that the system is able to profile users, notwithstanding the fact that in practice there is no information regarding the actual percentage of input categories. To better validate the approach, we are currently performing experiments considering user histories composed by more than two categories while studying the impact of the category imbalancement.

5. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, a system for user profiling able to identify user interests has been presented. The proposed system is able to profile users in terms of a set of given categories belonging to a taxonomy, together with a set of documents representing her/him. Experiments, performed using WordNet Domains as reference taxonomy and Wikipedia as document source, highlight that the proposed approach is quite effective. Due to the general-purpose machine learning techniques, we expect it to be exportable to other –more specific– taxonomies and to be used in the field of contextual advertising.

As for the future directions, we are studying how to improve the performance of the system. In particular, we are investigating new measures that may outperform the results obtained with TFIDF. Moreover, we are planning to adopt community detection algorithms to suitably prune the semantic net generated by the semantic words analyzer.

6. REFERENCES

[1] A. Addis, M. Angioni, G. Armano, R. Demontis, F. Tuveri, and E. Vargiu. A novel semantic approach to document collections. In *IADIS Multi Conference on Computer Science and Information Systems 2008*, 2008.

[2] M. Agelli, G. Armano, G. Cherchi, M. Clemente, and D. Ghironi. Experimenting combinations of content-based and collaborative filtering with a photo recommender system. *Communications of SIWN*, 5(August 2008):33–38, 2008.

[3] R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[4] C. Deepayan, A. Deepak, and J. Vanja. Contextual advertising by combining relevance with click feedback. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2008. ACM.

[5] A. Kobsa. User modeling: Recent work, prospects and hazards, 1993.

[6] L. Larkey. Automatic essay grading using text categorization techniques. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95, New York, NY, USA, 1998. ACM.

[7] G. Linden, B. Smith, and J. York. Amazon.com recommendations. *IEEE Internet Computing*, 07(1):76–80, 2003.

[8] B. Magnini and G. Cavaglià. Integrating subject field codes into wordnet. In *Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.)* Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, pages 1413–1418, 2000.

[9] S. E. Middleton, N. R. Shadbolt, and D. C. D. Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, 2004.

[10] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[11] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1984.

[12] B. Sarwar, G. Kaypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *IEEE Internet Computing, 10th International World Wide Web Conference*, pages 285–295, 2001.

[13] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–55, 2002.

[14] F. C. Stevens. *Knowledge-based assistance for accessing large, poorly structured information spaces*. PhD thesis, Boulder, CO, USA, 1993.

[15] T. W. Yan and H. Garcia-Molina. The SIFT information dissemination system. *ACM Transactions on Database Systems*, 24(4):529–565, 1999.