Identifying Inter-Domain Similarities through Content-Based Analysis of Hierarchical Web-Directories

Shlomo Berkovsky[1], Dan Goldwasser[1], Tsvi Kuflik[1] and Francesco Ricci[2]

Abstract. Providing accurate personalized information services to the users requires knowing their interests and needs, as defined by their User Models (UMs). Since the quality of the personalization depends on the richness of the UMs, services would benefit from enriching their UMs through importing and aggregating partial UMs built by other services from relatively similar domains. The obvious question is how to determine the similarity of domains? This paper proposes to compute inter-domain similarities by exploiting well-known Information Retrieval techniques for comparing textual contents of the Web-sites, classified under the domain nodes in Web-directories. Initial experiments validate feasibility of the proposed approach and raise open research questions.

1 INTRODUCTION

Providing accurate personalized services to the users requires modeling their preferences, interests and needs. This data is referred in the literature as the User Model (UM) [3]. Typically, service providers build and maintain proprietary UMs, tailored to the domain of the service and UMs representation, dictated by the personalization technique being exploited. Since the quality of the provided personalization heavily depends on a richness of the UMs, different services would benefit from enriching their UMs through importing and aggregating partial UMs, i.e., the UMs built by other, possibly related, services. This will bootstrap the UMs of the services where no UMs exist and enrich already existing UMs, leveraging the quality of the provided personalization [1].

One of the major issues that should be resolved to facilitate proper aggregation of partial UMs is "*Which services can provide valuable partial UMs?*". Considering a wide variety of application domains, we conjecture that in addition to the services from the same application domain, also the services from other (relatively similar) domains can provide valuable partial UMs. However, this inherently brings a question of "*Which application domains are considered as similar domains?*".

This work proposes an automatic approach for devising inter-domain similarities through content-based analysis of the Web-sites, classified under the application domains. Such classifications can be found in various human-edited Web-directories, e.g., Google Directory (*http://directory.google.com*), Yahoo! Directory (*http://dir.yahoo.com*), or Open Directory (*http://dmoz.org*). In Web-directories, the Web-sites are classified under application domains, which are recursively partitioned to more specific sub-domains, and so forth. Since Web-directories are edited manually by human experts, we assume that classification of the Web-sites under domains and sub-domains is correct and that large enough set of Web-sites, classified under a certain domain, can be considered as a reliable representative of the domain contents.

Hence, we base inter-domain similarity computation on a comparison of the textual contents of the Web-sites contained in each domain. For this, we first model the contents of the classified domain Web-sites using well-known Information Retrieval (IR) indexing technique of TF-IDF [4]. It allows representing a domain as a ranked vector of terms (derived from the classified Web-sites), where domain-specific terms are assigned higher weights. Then, inter-domain similarity is computed using a cosine similarity metric [4], where the similarity of two vectors is computed as the cosine of the angle between them in a multi-dimensional space. Initial experimental results, conducted over two Web-directories validate the feasibility of the proposed inter-domain similarity computation approach and raise open questions for future research.

The issue of computing similarity over a hierarchical domains structure is elaborately discussed in [2]. That paper presents, analyzes and experimentally compares a set of similarity metrics over a hierarchical tree of domains, while focusing on exploiting the hierarchical structure as an indicator for the similarity values. Conversely, in this work we aim at computing domains similarity through content-based analysis of the Web-sites, classified to the nodes of Web-directories, whereas hierarchical structure of the directories can serve as a heuristic limitation for the search process.

2 ANALYZING DIRECTORIES' CONTENTS

We suggest calculating inter-domain similarity basing on the textual contents of the Web-sites classified under the domain node in a Web-directory. The proposed approach is based on an inherent assumption that the textual contents reliably represent the domain. Although this assumption is not always true (in addition to the textual contents, nowadays Web-sites contain various graphical, audio and video objects, and use dynamic Web-technologies to generate their contents), we believe that large enough number of classified Web-sites will provide an accurate domain representation and a stable basis for the similarity computation.

To represent textual contents of the Web-sites classified under domains of a Web-directory, we propose to exploit well-known IR indexing techniques. For the indexing, stop-words (such as *and*, *to*, *the*, etc…) and HTML tags are filtered-out, the terms are lemmatized to their basic grammatical forms, and weighted according to their domain-related TF-IDF values, assigning higher weights to domain-specific terms (i.e., terms that are frequent in the domain Web-sites only) [4]. As a result, a ranked vector of weighted domain terms, representing the contents of the Web-sites under a given application domain, is obtained. For example, consider a vector of the top terms in *football news* domain, shown in Table *1*. It can be clearly seen that the list reliably represents football terms.

Table 1. Top-ranked terms in football news domain

term	score	high	senator	gameface	rumor	league	archive
weight	462.56	397.14	353.94	320.38	251.88	128.31	103.16

After the domain representative vectors are constructed, inter-domain similarity can be easily computed using one of the existing similarity metrics. In this work, we used cosine similarity metrics, defining the similarity between two vectors as the cosine of the angle between them in a multi-dimensional space:

[1] University of Haifa, Haifa, Israel
[2] ITC-irst, Trento, Italy

$$sim(V_x, V_y) = \frac{V_x \bullet V_y}{\| V_x \|_2 \times \| V_y \|_2}$$

where $\bullet$ denotes the dot product between the vectors, and $\|V_i\|_2$ denotes the *2*-norm of the vector (i.e., the square root of the sum of the squares of the vector elements). The result of the cosine similarity computation is a single scalar, reflecting the similarity of the respective domains, based on their textual contents.

To obtain the required inter-domain similarity values, the above process of similarity computations should be repeated for all the possible pairs of domains. However, compound structure of nowadays Web-directories and high number of the existing domains, make this task expensive and pose a need for heuristically limiting the search space. We propose to exploit the inherent hierarchical structure of the Web-directories for this purpose. Since the Web-directories are organized as linked hierarchical structures, the search (and similarity computations) may be heuristically limited to a set of relatively close nodes only. For example, a search for the domains, similar to *Arts→Movies→Genres→Drama*, may be limited to other movies genres *Arts→Movies→Genres→**, other sub-domains of movies *Arts→Movies→**, and more abstract higher-level domain of visual arts *Art→Visual Arts→**.

3 EVALUATION AND FUTURE RESEARCH

We have implemented a prototype system for inter-domain similarity computations. To obtain the contents of the domain Web-sites, we exploited a simple Web-crawler, configured to download the sites' textual contents only. The contents of the domains were represented using IR indexing techniques employed by Lucene, open-source search engine (*http://lucene.apache.org*). The experiments were conducted over Google Directory and Open Directory.

The first experiment was designed to validate the proposed approach for devising inter-domain similarity values. For this, we downloaded the contents of a small set of application domains and computed their pair-wise inter-domain similarities using the above two Web-directories. The pairs of domains whose similarity was computed were: (a) *Arts→Television→News→News on the Internet* vs. *Arts→Television→News→Sports* (b) *Arts→Television→News* vs. *Sports→Football→NCAA*, and (c) *Computers→Hardware→Storage* vs. *Sports→Football→NCAA*. Note that these domains were selected as their similarity classification can be easily done by a common sense. Pair (a) is a relatively similar one, since the nodes are siblings and many news Web-sites have a dedicated sports section. Pair (b) is similar to some extent only. Despite the fact that the nodes are distinct in Web-directories, news domain does contain some NCAA news items, however, it contains also many other news topics. Pair (c) is highly dissimilar, as the terminology used in hardware storage devices' domain and in NCAA football is very different. Table *2* summarizes the experimental results.

Table 2. Inter-domain similarities in different Web-directories

	(a)	**(b)**	**(c)**
Google Dir.	0.3826	0.2676	0.1416
Open Dir.	0.4205	0.1879	0.1103

Experimental results validate feasibility of the proposed approach, as the similarity values for pair (a) are higher than for pair (b), which, in turn, is higher than for pair (c). However, the results raise a question regarding the stability of the proposed approach with the number of the Web-sites that are indexed. Both Google Directory and Open Directory are structured as highly connected graphs, where each domain comprises a set of sub-domains, whereas the sub-domains of different domains may be interlinked. Since a Web-site may be classified to one of the sub-domains of a given domain, the second experiment was designed to check the impact of indexing and computing TF-IDF representation of a larger set of Web-sites, rather than the sites classified directly under the node of a given domain. We compared three expansion policies: (i) considering also the Web-sites classified under the sub-domain nodes of a given domain, (ii) considering also the Web-sites pointed by the outgoing links from the Web-sites classified under the node of a given domain, i.e., increasing the depth of the search on the Web, and not in the Web-directory, and (iii) integrating (i) and (ii). These policies were employed on the above three pairs of domains and the similarity computations were repeated for a larger set of Web-sites. Table *3* summarizes the results.

Table 3. Inter-domain similarities using different expansion policies

		(a)	**(b)**	**(c)**
Google Dir.	**(i)**	0.3555	0.4279	0.2709
	(ii)	0.3632	0.2275	0.1829
	(iii)	0.4189	0.3874	0.2888
Open Dir.	**(i)**	0.2703	0.4929	0.1045
	(ii)	0.3354	0.2164	0.0841
	(iii)	0.3279	0.3372	0.1232

The impact of policy (i) can be defined as negative, as the similarity of the of similar domains in pair (a) decreases, while the similarity of less similar pair (b) and (c) increases. We hypothesize that this can be explained by the fact that indexing the Web-sites of the sub-domain nodes inserts a decent noise to the TF-IDF vectors, as also less similar Web-sites are indexed and graph structure of Web-directories brings in sub-domains, which are actually sub-domains of other dissimilar domains. The same observation is true also for policy (ii) as the similarity of (a) decreases, while the similarity of (b) and (c) increases. However, the impact of (ii) is significantly weaker, and (a) is still more similar than (b), which is more similar than (c). This allows us to conclude that indexing the Web-sites pointed by the outgoing links inserts less noise than indexing the Web-sites classified under the sub-domains of a given domain. The impact of (iii) is unclear and it depends on the specific application domain and additional Web-sites that are indexed.

Although the experiments initially validate the feasibility of the proposed inter-domain similarity computation approach, elaborate experiments and analysis are still required. In the future, we plan to analyze statistical properties of the domains for concluding regarding the conditions and domains, where the proposed approach is applicable. We plan to study the use of the graph structure of Web-directories for heuristically limiting the similar domains search. For this, we will check the correlation between the vicinity of nodes in the Web-directories' graph and their content-based similarity. Also, we plan to check the feasibility of the above heuristics in sparse application domains with a small number of classified Web-sites.

As these questions are answered, we plan to investigate the task of aggregating the partial UMs. We believe that the proposed similarity computation approach can be a stable basis for a dynamic user modeling mechanism, facilitating efficient provision of accurate personalization services.

REFERENCES

[1] S.Berkovsky, "Ubiquitous User Modeling in Recommender Systems", in proceedings of the UM Conference, Edinburgh, 2005.

[2] P.Ganesan, H.Garcia-Molina, J.Widom, "Exploiting Hierarchical Domain Structure to Compute Similarity", in ACM Transactions on Information Systems, vol.21(1), pp.64-93, 2003.

[3] A.Kobsa, "Generic User Modeling Systems", in User Modeling and User-Adapted Interaction, vol. 11(1-2), pp.49-63, 2001.

[4] G.Salton, M.McGill, "Introduction to Modern Information Retrieval", McGraw-Hill Publishing, 1983.