Profiling Web Users Preferences with Text Mining

Plan2See – dynamic formulation of conceptual groups

Pedro Bonifácio Costa
ISCTE-IUL
Lisboa, Portugal
pedro.bonifacio.costa@gmail.com

Sancho Oliveira
ISCTE-IUL
Lisboa, Portugal
sancho.oliveira@iscte.pt

Luís Nunes
ISCTE-IUL
Lisboa, Portugal
luis.nunes@iscte.pt

Abstract — **In this paper we propose a new approach to clustering Web-based content that could be leveraged by users' preferences. These preferences may imply grouping or dividing the initial clusters so that the resulting clusters represent users' profiles. This approach could be applied to recommend different textual contents, based on the clusters or per user. We conduct a summarized survey of what is the state-of-the-art towards a personalized and interactive manipulation of data by users and present the general configuration of our method and some preliminary results well as the objectives for future work.**

Text Mining; Clustering; Profiling; Statistical Models; Web 2.0

I. Introduction

The research in recommender systems has shown that there is still space to explore new systems and improve the existing ones. We're conducting the research on a "hybrid" approach, such as in [1] – authors mention three categories: content-based, only the content of an item is relevant to any recommendation; collaborative, users' interaction is what builds recommendations; hybrid, the recommendations are mixed from the analysis of items and users' actions – because the content of items being recommended has the same relevance as the historical preferences shown by users.

What we intend is to build a dynamical, adaptive and feasible method to extend users' profiles using information gathered from other users.

The clustering of contents is complemented with information of users' preferences and this determines the new grouping or division of each cluster of documents where circles of interests are built around users' choices providing more insight of what users prefer.

This technique is applied to location-based events where experimental users are invited to select or unselect those events that are relevant to them. The choice is not restrict to location or time constraints as only the name and the description is shown.

The resulting clusters could be applied to other textual contents to conduct customer profiling or shopping cart such as in [2].

The convergence of the clusters, resulting from the application of KMeans algorithm to our dataset, with the users' profiles, acquired with their selection of items, is one of the proof-of-concepts we'll use.

To test our results we'll apply acquired profiles and clusters to a different dataset (e.g.: events versus news articles) within the same lexicon (e.g.: information technology).

II. Related Work

In unstructured textual documents one can use machine learning techniques, like classification or clustering, to conduct data analysis [3] and build recommender systems, based on textual items. These systems are capable of relate and recommend new items to users using their profiles to determine which new items are the most appropriate recommendations.

Application of the term-based frequencies vectors to represent each document as in [4] has been the most common way to conduct comparisons and find patterns in textual methods where frequencies or binary occurrences may be discarded in favor of Term Frequency – Inverse Document Frequency (TF-IDF) transformation [5]. The potential growth of such intelligent applications is enormous and thus can deliver new deals and leverage current businesses feasibility and profitability [6].

A. Documents Categorization

To determine the best approach we've analyzed forms of categorizing documents in groups found in the literature related to Data Mining, more specifically classification algorithms, such as those in [7] where we intended to use all historical behavior from such users.

The Memory-Based Approach and the Naïve Bayes Classification [8] are among the most common ways to categorize new emails as spam or legitimate and later works try to apply the same techniques to users' profiles [9].

In [10] authors applied Latent Semantic Analysis to transform narrow terms in more broad and meaningful terms to represent each group of documents, clustered by unsupervised learning.

B. Profiling Patterns

Recommender systems may require initial profiles [11] as the classification mechanisms are not trained for each user's preferences.

Also in [13] tagging mechanisms form the basis to evaluate recommendations where relevance presented by each

freepaper.me

object is combined with distance measures to form the final recommender system.

Although these approaches can be more common that the one in [14] the latter takes in account other information that implicitly profiles each user by tagging each user to the website that has been visited and thus analyze which items were related to that user's profile. The profiles are made through the neighborhood of related websites, based on each website terms.

Recommender systems may also demand undesired effort for users' to evaluate if the new items are appropriate [13].

In [15] authors mention the subjectivity of users' input to construct profiles based in content tagging and that this construction requires deeper knowledge from users'. Although their experiment is somehow related to ours, as it builds profiles from real-time usage and aggregates information between users, but it uses access logs to personalize contents of a unique website – it does not extends profiles to other datasets.

These experiments in a unique source could be somehow better for large enterprises in which each enterprise website has plenty of information and users' itself. The research in [16] also builds profiles from access logs but the major difference is in gathering information from users automatically as it tracks the evolving profiles through several web sessions.

For our goal we intend not only to build profiles with multiple purposes but that the system could also recommend users' from different topics – represented by groups of concepts. In [17] this goal is achieved with classification mechanism, where each user must present his profile prior to the recommendation which requires the undesired effort as mentioned above. Our in [18] where an "explicit" profile is built prior to the redefinition of web search mechanisms with the same goal of personalize web search results.

The benefits of building empirically users' profiles without prior interaction is also mentioned in [19] where web search results are redefined towards personalization. The authors conducted experiments to use users' context, i.e. the historical behavior and related users' profiles, to achieve better recommendations – where the evaluation of the mechanism was based on Relevance Feedback query reformulation strategy – which could be of use to our present work if we were basing the recommendations on users' input of desired items. In [20] authors also intend to improve web search as conducting an analysis on how the grouping of items into concepts and theirs' relations could leverage personalization of resulting items in a web search query. The same goal is mentioned in [21] where users' behaviors is analyzed to build profiles that are then matched with pre-analyzed web search results to a given query, where profiles and contents are described according to the same ontological references.

III. Dynamic Formulation Of Conceptual Groups

In our approach a new method is proposed that leverages KMeans flexibility and performance in content-clustering with awareness of users' profiles. This approach allows clusters to acquire subjective meaning as users' actions imply transformations in KMeans initial clusters.

A. Collecting Resources

Our first requirement was to have items in which users could explore theirs interests and choose the relevant for them.

To accomplish our goals we've decided that the items were events – with a title, location, description and date – that users could select planning to attend.

Our main goal was simply to test this approach. We've determined to use Information Systems (IT), generally speaking, related events. This decision was partly because test-users with that profile were available for our usability experiments.

In order to achieve a large enough data set we've built a Web crawler to navigate automatically through predetermined websites, repositories of IT events, and extract all events in those repositories to our own database.

B. Content Analysis

When conducting the content analysis we start by comparing Weka[1] and RapidMiner[2] and we've decided to use RapidMiner, given the analysis summarized in Tab. I.

TABLE I. Machine Learning Systems Comparison

Alternative	Open Source License	Built-in machine learning	Final Process Drag-and-drop Design	Built-in stemmers and lemmatizers	Built-in API
Weka	Yes	Yes	No	No	Yes
RapidMiner	Yes	Yes	Yes	Yes	Yes

The database contains 4194 events (our universe or documents). There are around 6700 unique terms not considering punctuation or other common English stop words. The 6700 terms considered only the base form of stemmed words – reducing words to their syntactical base.

C. RapidMiner Process

Our process retrieves all events universe and for each one tokenizes its terms and reduces terms to their base form. The terms universe are reduced by top and bottom limits of occurrences and the clustering by Kmeans takes place.

After debugging we've concluded that most of the terms were not semantically significant and processes ran on RapidMiner led us to conclude that around 150 terms (if some terms occurs below 20 percent and above 0.2 percents of the universe it is used) and 10 initial clusters was more appropriate because: 1) above 150 terms there were only 2 referring documents which may imply high interrelated concepts; 2) above 10 clusters these start to be almost "empty" with only 2 or even 1 events; and 3) below 10 clusters some of them have almost 30% of the universe, what also may imply the same as in 1).

These results are consistent with our expectations for

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://rapid-i.com/content/view/181/190/

freepaper.me

events selected only in the IT universe.

D. Initial Clustering

The initial 10 clusters are calculated by KMeans algorithm based on absolute frequencies of each term on the dictionary in each event.

In our approach the documents are gathered and clustered before the initial deployment of the application to users to guarantee that there is already a recommendation based only in the content. The RapidMiner process deploys 10 clusters and the terms' absolute frequencies on application startup.

The application deployment takes care of gathering events', frequencies of terms – based on a dictionary of 150 words – and acquires the events' groups based on the definition files deployed by RapidMiner.

E. Users' selection of events

In Plan2See users' selection of events impacts the initial data structure that may be transformed to respond to these selections.

An interface was developed with aid of the Tapestry framework where users can select and unselect events and also search for relevant subjects.

Whenever a selection is made, whether an event is selected or unselected, the application has to evaluate if changes are to be made in the existing clusters, by grouping similar users' interests or dividing groups to narrow these interests.

F. Grouping and Dividing

Every selection is evaluated by its' importance to the data model. If users often select events from any two ungrouped clusters then the application should evaluate the possibility of grouping the clusters. If the application detects that users are selecting only part of a cluster and that part is not related to the remaining events in that cluster the application should consider dividing this cluster.

These changes happen for all users' selections but occur for each user event to simplify our usability tests.

When any cluster is divided two new clusters arise based on the old one. The dividing mechanism selects each event in the old cluster and computes the Euclidean Distance from the event to the centroid of the user selected events (*a*) and the Euclidean Distance from the event to the centroid of the old cluster without the selected events (*b*) – if *a* is smaller than *b* the event goes to a cluster; otherwise it goes to the other.

G. Evalutian Relation Between Subsets

Above we've presented the method to determine if grouping or dividing is required based on subsets relation evaluating. To calculate if two subsets of events are related or unrelated we conduct a T-Test on terms frequencies of both subsets.

The steps to conduct a T-Test are:

I. Calculate the mean points for each subset, based on terms frequencies as "centroids";

II. Discard irrelevant dictionary term *t* that do not occur in any document *I* on both subsets;

$$\exists\, frequency(t_i) > 0 \tag{1}$$

III. A list of relevant terms is built upon a percentage, which is a configurable variable, of the terms with the higher frequencies in both subsets which results in k;

IV. Set *k* as the number of relevant terms;

V. Calculate each term's frequency f_t and standard deviation σ_t;

VI. Calculate confidence interval where *T* represents the cumulative probability of a T-Distribution with *k* degrees of freedom

a. Calculate confidence interval bottom value b_t for each term as in Equation (2).

$$b_t = \frac{f_t - T * \sigma_t}{k} \tag{2}$$

b. Calculate confidence interval top value v_t for each term as in Equation (3).

$$v_t = \frac{f_t + T * \sigma_t}{k} \tag{3}$$

VII. If the b_t values of a subset are between the confidence interval of the other subset or if the v_t values of a subset are between the confidence interval of the other subset then the subsets are related and grouping of subsets can be done; otherwise dividing a cluster can be made.

IV. Preliminary Results

We've conducted three tests to verify the application capability with the previous logic. In these tests parameters were set so that a small number of user actions could cause changes in the clustering. In a real system more evidence should be gathered before actual changes could occur.

In the first test one group was built were the main subject was not precisely defined. The events' categories ranged from *AGILE* project management or *CLOUD COMPUTING* as well as other subjects. While selecting events, with only one user, focused only in a specific category, such as only selecting events with *AGILE* on the description and title, the application started dividing the cluster and a new group based on *AGILE*-related documents was formed.

Test two took place with two well-defined groups: one group of events related to *SCRUM* and other related to *AGILE*. While selecting events, with only one user, focused in both groups the application started grouping both clusters in one.

The third test had the purpose of grouping and dividing clusters with more than one user. We've built two clusters each one with its category: *SCRUM* versus *AGILE*. Both clusters had the general category *PROJECT*. User 1 selected some events from both groups and the application grouped all. When user 2 selected some events only with AGILE the application divided the clusters and the original datasets were reestablished.

We've concluded that the application is now ready to be used in live mode with several users selecting events' preferences.

freepaper.me

V. FUTURE WORK

Although this research is under progress we've already determined the form to conduct our resulting clusters analysis as we'll collect real users' preferences with the application deployment.

While collecting users' preferences we'll validate if the dictionary, i.e. the terms that are used for distances measures, are appropriate. Several experiments with different dictionaries will have to be conducted to determine which one is the best fit for our dataset.

The clusters resulting from the application deployment need to be checked against users' preferences. The objective is to check if clusters' centroids are moving towards users' preferences using measures like Euclidean Distance. This evaluation depends on the final results of the application deployment.

The effectiveness of profiles most be certified by evaluating results with other content items of the same type. We will use other content, like IT news articles, as new items to recommend when evaluating clusters' feasibility. For nw items that correspond to any cluster, given its content, we'll present users the ones that are relevant (if the user has selected any event from that cluster) – positive or negative feedback will be asked to evaluate results.

REFERENCES

[1] Adomavicius, G. & Tuzhilin, A. "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE Transactions on Knowledge and Data Engineering, vol. 17, pp. 734-749, June 2005.

[2] Chidanand, A. et al. Business Applications of Data Mining. Communications of the ACM - Evolving data mining into solutions for insights. Volume 45 Issue 8 [online] August 2002 [cited 1 Apr 2013]. Available from Internet: http://dl.acm.org/citation.cfm?id=545178

[3] Kroeze, Jan H.; Matthee, Machdel C.; Bothma, Theo J.D. Differentiating Data- and Text-Mining Terminology. Proceeding SAICSIT '03 Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology [online] 2003 [cited 24 Apr 2012]. Available from Internet: http://dl.acm.org/citation.cfm?id=954024 ISBN:1-58113-774-5

[4] Weiss, M. Sholom et al. Text Mining: Predictive Methods for Analyzing Unstructured Information. 1st Ed. New York. Springer. 2004.

[5] Leopold, E.; Kindermann, J. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? Machine Learning Volume 46 Numbers 1-3 [online] January 2002 [cited 1 May 2012] Available from Internet: http://www.springerlink.com/content/q2841573463j066p/fulltext.pdf

[6] Witten, I. H.; Eibe, F.; Hall, M. A. Data Mining: Practical Machine Learning Tools and Techniques 3rd edition. Burlington, MA : Morgan Kaufmann, 2011. ISBN 978-0-12374856-0

[7] Costa, P.. Plan2See – discovering events through Web 2.0 [online] July 2011. [cited 25 Apr 2012]. Available from Internet: http://dl.acm.org/citation.cfm?id=2016727

[8] Androutsopoulos et at. Learning to Filter Spam E-mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach. [online] September 2000. [cited 25 Apr 2012]. Available from Internet: http://arxiv.org/ftp/cs/papers/0009/0009009.pdf

[9] Lakshmi, K. & Mukherjee, S. Profile Extraction from Mean Profile for Automatic Text Categorization. [online] 2005. [cited 25 Apr 2012]. Available from Internet: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1631499&userType=inst

[10] Lourenço, A. et al. Unsupervised Organization of Scientific Documents. [online] October 2011. [cited 1 May 2012] Available from Internet: http://www.it.pt/papconf_abs_p.asp?ID_PaperConference=10736&id=3

[11] Pazzani, M; Billsus, D. Learning and Revising User Profiles: The Identification of Interesting Web Sites [online] June 1997 [cited 1 May 2012] Available from Internet: http://dl.acm.org/citation.cfm?id=261098

[12] Diederich, J.; Iofciu, T.; Finding Communities of Practice from User Profiles Based on Folksonomies [online] 2006 [cited 1 May 2012] Available from Internet: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.1617&rep=rep1&type=pdf

[13] Sugiyama, K; Hatano, K.; Yoshikawa , M. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users [online] 2004 [cited 1 May 2012] Available from Internet: http://dl.acm.org/citation.cfm?id=988764

[14] Hoebel, N. et al; The Design of Gugubarra 2.0: A Tool for Building and Managing Profiles of Web Users [online] 2006 [cited 1 August 2012] Available from Internet: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=4061385&contentType=Conference+Publications

[15] Mobasher, B. et al; Discovery of Aggregate Usage Profiles for Web Personalization [online] 2002 [cited 1 Apr 2013] Available from Internet: http://link.springer.com/article/10.1023/A%3A1013232803866

[16] Soliman, M. et al; A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites [online] 2008 [cited 1 Apr 2013] Available from Internet: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4358953&tag=1

[17] Pazzani, M.; Billsus, D.; Learning and Revising User Profiles: The Identification of Interesting Web Sites [online] 1997 [cited 1 Apr 2013] Available from Internet: http://link.springer.com/article/10.1023/A%3A1007369909943

[18] Anane, R.; Odetayo, M.; James, A.; Explicit User Profiles in Web Search Personalisation [online] 2011 [cited 1 Apr 2013] Available from Internet: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5960107

[19] Sugiyama, K.; Hatano, K.; Yoshikawa, M.; Adaptive Web Search Based on User Profile Constructed without Any Effort from Users [online] 2004 [cited 1 Apr 2013] Available from Internet: http://dl.acm.org/citation.cfm?id=988764

[20] Sieg. A.; Mobasher, B.; Burke, R.; Web Search Personalization with Ontological User Profiles http://dl.acm.org/citation.cfm?id=1321515

[21] Sieg. A.; Mobasher, B.; Burke, R.; Ontological User Profiles for Representing Context in Web Search [online] 2007 [cited 1 Apr 2013] Available from Internet: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4427547

freepaper.me