Multi-evidence User Group Discovery in Professional Image Search

Theodora Tsikrika[1] and Christos Diou[2]

[1]Information Technologies Institute, CERTH, Thessaloniki, Greece
[2]ECE Department, Aristotle University of Thessaloniki, Thessaloniki, Greece
theodora.tsikrika@iti.gr, diou@mug.ee.auth.gr

Abstract. This work evaluates the combination of multiple evidence for discovering groups of users with similar interests. User groups are created by analysing the search logs recorded for a sample of 149 users of a professional image search engine in conjunction with the textual and visual features of the clicked images, and evaluated by exploiting their topical classification. The results indicate that the discovered user groups are meaningful and that combining textual and visual features improves the homogeneity of the user groups compared to each individual feature.

1 Motivation

The discovery of groups of similar users is useful to information retrieval applications, such as ranking and sponsored search, that benefit from adapting their results to users' interests. Given that data sparseness and ambiguity may lead to ineffective representations of user interests, especially for users with limited history, research has suggested to leverage evidence obtained from the user group(s) the individual belongs, to either augment their personal profile, and thus potentially lead to more effective personalisation, or to use such evidence for 'groupisation', i.e., adapt retrieval the same way for all group members [8].

Previous studies have explored several features for discovering user groups, including occupational and topical evidence [8], demographic information [8, 5] with a particular focus on gender [6], geographical context [1], and reading level efficiency [2]. Their results show the usefulness of such evidence for improving retrieval. Additional studies [10, 11] have investigated user segmentation outside the context of a particular retrieval application with the goal to gain an understanding of how the search behaviour of specific groups differs. To this end, they analysed large scale web search logs in terms of the users' query topics and/or session characteristics, together with the users' demographic profile information augmented with U.S. census data. Their results showed that it is possible to identify distinct patterns of behaviour along different demographic features.

Our study aims at further developing an understanding of user group identification by investigating the hypothesis that the combination of multiple evidence improves the discovery of users groups with similar topical interests compared to the groupings based on individual features. To this end, we analysed the search

log data collected by the commercial picture portal of a European news agency for a sample of 149 of their registered users, in conjuction with the captions, visual features, and a topical classification of the available images. Our analysis takes place in a context that is different to the search environments examined in all of the above work in at least one of the following aspects. Ours is a professional, rather than a web, environment, and furthermore it is oriented towards image, rather than text retrieval. Also, although we examine topical features for discovering user groups, similarly to [8], we consider the users' log activity rather than data collected through a user study. Moreover, to the best of our knowledge, visual features have not been previously investigated in such a context.

2 Data Acquisition and Processing

The search log data used in this study were collected by the news agency over a two year period (June 2007 – July 2009), with a three-month hiatus (October – December 2007). The sample considered in this study consists only of registered users logged into their account and was processed as follows. First, the logs were segmented into sessions, i.e., series of a single user's consecutive search actions assumed to correspond to a single information need. No intent-aware session detection was applied [3], and session boundaries were identified when the period of inactivity between two successive actions exceeded a 30-minute timeout, similarly to [11]. Next, the queries were 'lightly' normalised by converting them to lower case and removing punctuation, quotes, special characters, extraneous whitespace, URLs, and the names of major photo agencies. Furthermore, empty queries and queries consisting only of numbers or whitespace characters were removed. No stemming or stopword removal was applied at this stage. Furthermore, consecutive identical queries submitted in the same session were conflated. The final step was to further sample the logs so as to include only "active" users, i.e., those who had issued at least 10 queries, with each followed by at least one

Table 1. IPTC subject codes and the IPTC distribution of clicked images.

	IPTC subject codes		#images	% classified images	% all images
	Code	Name			
1.	ACE	Arts, Culture, & Entertainment	36,413	14.3%	12.5%
2.	CLJ	Crime, Law & Justice	9,768	3.8%	3.4%
3.	DIS	Disaster & Accident	6,284	2.5%	2.2%
4.	EBF	Economy, Business & Finance	16,148	6.4%	5.6%
5.	EDU	Education	599	0.2%	0.2%
6.	ENV	Environmental issue	1,819	0.7%	0.6%
7.	HTH	Health	2,014	0.8%	0.7%
8.	HUM	Human interest	12,528	4.9%	4.3%
9.	LAB	Labour	2,006	0.8%	0.7%
10.	LIF	Lifestyle & Leisure	2,411	0.9%	0.8%
11.	POL	Politics	39,473	15.5%	13.6%
12.	REL	Religion	2,111	0.8%	0.7%
13.	SCI	Science & Technology	1,708	0.7%	0.6%
14.	SOI	Social issue	1,385	0.6%	0.5%
15.	SPO	Sport	110,090	43.3%	37.9%
16.	WAR	Unrest, Conflicts, & War	7,608	3.0%	2.6%
17.	WEA	Weather	2,093	0.8%	0.7%
			254,458	100%	87.6%

click. Our final sample thus contains 149 users who submitted a total of 113,176 queries (55,386 unique) and clicked on a total of 519,849 images (290,593 unique).

In addition to textual captions, the content of images is also described using the top level of the International Press Telecommunications Council's (IPTC) hierarchical newscodes, i.e., the 17 IPTC *subject* codes (`http://cv.iptc.org/newscodes/subjectcode/`) listed in Table 1. Out of the 290,593 unique images clicked in our sample, 254,458 (87.6%) had been manually classified by the news agency's archivists. Table 1 lists the distribution of the clicked images over the 17 IPTC subjects. Sports is by far the dominant category, indicating a slight bias in the topical interests of the sampled users, as these are reflected by their searching behaviour. Politics and cultural topics follow in almost equal measures. Economics, human interest topics, crime, and war are the next subjects of interest in descending order, with the rest following in much lower percentages.

3 User Groups

Groups of users with shared topical interests are *implicitly* formed based on the hypothesis that such users issue similar queries and/or click on similar images. To this end, clustering is performed by employing user distances that exploit textual and visual evidence associated with the users' queries and/or clicks.

The first user distance, denoted as **text**, is the Euclidean distance between term vectors that correspond to user queries and clicked image captions. Initially the text of each user's queries and the captions of clicked images are concatenated. A term vector is generated by applying stemming, but without removing stopwords, and by estimating the term weights using a *tf.idf* scheme with normalisation, using the Text to Matrix Generator (TMG) Matlab toolbox [12].

The second, denoted as **vis**, is the visual distance of the users' clicked images, defined as follows. A vector of 4,096 features was extracted from each image using the 'bag of visual words' method, with dense keypoint sampling and the color SIFT descriptor [9]. Each user u is represented by a set of clicked images $\mathbf{I}_u$ and the distance between two users u, v is computed via the Hausdorff metric [4]:

$$d_H(u,v) = \max\left\{\max_{\mathbf{x}\in\mathbf{I}_u}\min_{\mathbf{y}\in\mathbf{I}_v} d(\mathbf{x},\mathbf{y}), \max_{\mathbf{y}\in\mathbf{I}_v}\min_{\mathbf{x}\in\mathbf{I}_u} d(\mathbf{x},\mathbf{y})\right\} \quad (1)$$

where $d(\mathbf{x},\mathbf{y}) = \frac{1}{2}\sum_i \frac{(\mathbf{x}(i)-\mathbf{y}(i))^2}{(\mathbf{x}(i)+\mathbf{y}(i))}$ is the χ^2 distance between two visual word histograms. The distance d_H is the maximum distance found between each image of $\mathbf{I}_u$ and its closest image in $\mathbf{I}_v$. It is often used to compare sets, such as sets of points, in computer vision or computer graphics applications; here, it produces a distance between the sets of feature vectors of the images clicked by u and v.

A composite distance is then formed by the weighted sum of the two criteria, after normalising *text* and *vis*: $d_u(u,v) = w_1 d_{\mathbf{text}}(u,v) + w_2 d_{\mathbf{vis}}(u,v)$. We investigate all possible combinations of $w_i \in [0,1]$ at step 0.1, such that $\sum w_i = 1$.

User groupings are generated by applying agglomerative clustering with these distances and several linkage criteria (*single*, *complete*, *average*, and *weighted*), and by also varying the number of non-overlapping clusters k, as discussed next.

4 Analysis

The above **text**, **vis**, and **text_vis** (w_1/w_2) groupings are analysed so as to investigate how meaningful they are, i.e., whether group members are more similar to each other with respect to their topical interests, than to members of other groups. In particular, the inter– and intra–cluster variation in people's topical interests is examined using the following inter– and intra–cluster pairwise proximity measures: *cohesion* and *separation* [7]. Cluster cohesion is defined as the average of the pairwise proximity values of all points within the cluster, while separation as the average of the pairwise proximity values of each point within the cluster to all points in all other clusters.

In the absence of explicit ground truth, the images clicked by users are considered as implicit indicators of their topical interests. The IPTC subject distribution of the images clicked by each user can then be considered as an implicit user profile that can be used for evaluating user groupings. Cohesion and separation are then estimated by employing the *Jensen-Shannon (JS) divergence* between such distributions, and in particular its value subtracted from 1, as a pairwise user proximity measure.

The use of a ground truth dataset based on the 17 IPTC subjects motivates us to focus our analysis on clusterings with $k = 17$. Given, though, that 9 IPTC subjects are associated with less than 1% of all clicks each, it is highly likely that the topical interests of the users in our sample gravitate towards the most dominant 8 IPTC subjects. Therefore, our analysis varies k between 8 and 17.

Figure 1 shows the cohesion and separation for the groupings generated by the *weighted* linkage agglomerative clustering, which produces the best results among the different linkage criteria. For each k, *text* has higher cohesion than the whole sample, since it improves over the *average cohesion* computed over all possible user pairs. It also produces more cohesive groups than *vis*. Whereas *vis* is a rather weak source of evidence, its combination with *text* using 0.9/0.1 as weights improves significantly the cohesiveness of the groups against those generated by each individual feature. Weighing higher the contribution of *vis* is nevertheless detrimental for all other combination weights, as shown in Figure 1 for weights 0.8/0.2; the results for the remaining combinations are not shown.

Figure 1 further indicates that *text* produces user clusters that are better separated from the other clusters in the grouping, compared to the separation among clusters observed in the other groupings (the lower the separation value, the better). However, the differences in the separation values between *text* and *text_vis* (0.9/0.1) are rather insignificant, particularly for $k \geq 12$ where they are less than 0.02 (3%). Furthermore, a comparison of the cohesion values against the separation values for each of the *text* and *text_vis* groupings indicates that groups members are more similar to each user than to users in other groups. Therefore, our results indicate that combining textual and visual features for discovering user groups can significantly improve the homogeneity of the formed groups compared to the groupings generated by each individual feature.

To provide further support to our hypothesis, our user groupings are evaluated against an additional implicit ground truth dataset: a user classification

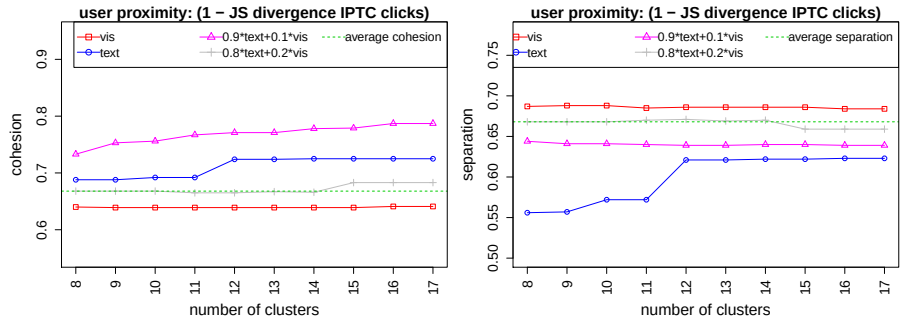

Fig. 1. Variation in group membership by comparing each grouping's intra–cluster cohesion and inter–cluster separation to the average similarity of all user pairs.

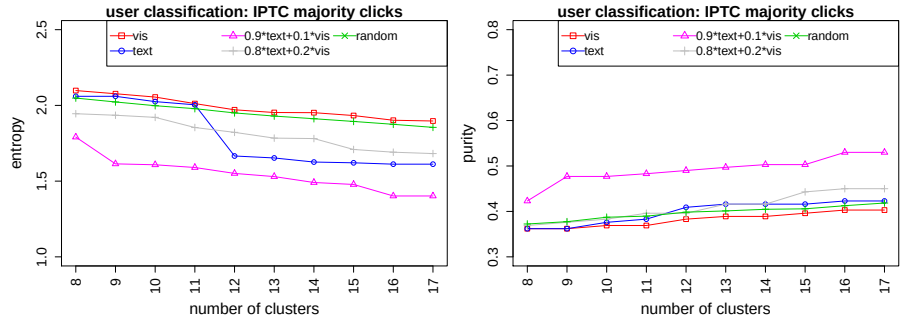

Fig. 2. Entropy and purity with respect to the *IPTC majority clicks* ground truth.

formed by assigning to each user the IPTC subject code of the majority of their clicked images. This corresponds to the 8 groups listed in Table 2. Given that, for personalisation purposes, it is more important to have groups that are more cohesive than well separated, our analysis focusses on measures that evaluate the extent to which a cluster contains objects of a single class: (i) the *entropy* of a cluster over the class distribution in the ground truth, and (ii) its *purity*, i.e., the frequency of the most frequent class of the ground truth in a cluster [7]. The entropy (purity) of a clustering is computed as the sum of the entropy (purity) values of all clusters, each weighted by its size. Figure 2 shows that the *text_vis* (0.9/0.1) clustering produces groups with the lowest entropy and highest purity compared to the individual features and the other combinations.

Similar results are also observed for the *average* and *complete* linkage criteria, but with less pronounced differences among the groupings, whereas for the *single* linkage criterion, the *text* and *text_vis* groupings perform almost equivalently for

Table 2. User classification based on the IPTC subject of the majority of their clicked images. All IPTC subjects that represent less than 1% of the clicked images in Table 1, apart from REL, are no longer present. WAR is also missing.

IPTC	ACE	CLJ	DIS	EBF	HUM	POL	REL	SPO	Total
# users	39	2	1	9	12	37	1	48	149
% users	26.2%	1.3%	0.7%	6.0%	8.1%	24.8%	0.7%	32.2%	100%

all combination weights. Overall, our results indicate that the combination of visual and text features for discovering user groups can significantly improve the homogeneity of the formed groups, as measured by cohesion, entropy and purity.

5 Conclusions

Our results show that user grouping based on the combination of textual and visual evidence leads to the discovery of more cohesive groups compared to those formed using individual features, a finding that could benefit 'groupisation' or personalisation approaches with group-augmented user profiles. Future work will follow a number of directions, including the exploitation of session information, and the incorporation of such groups in personalisation approaches.

6 Acknowledgements

This work was supported by the MULTISENSOR project, partially funded by the European Commission, under contract number FP7-610411. The authors are also grateful to Belga press agency for providing the datasets used in this work.

References

1. P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In W. Y. Ma, J. Y. Nie, R. A. Baeza-Yates, T. S. Chua, and W. B. Croft, editors, *SIGIR 2011*, pages 135–144. ACM, 2011.
2. K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In C. Macdonald, I. Ounis, and I. Ruthven, editors, *CIKM 2011*, pages 403–412. ACM, 2011.
3. D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12):1822–1843, 2009.
4. J. Henrikson. Completeness and total boundedness of the Hausdorff metric. *MIT Undergraduate Journal of Mathematics*, 1:69–80, 1999.
5. E. Kharitonov and P. Serdyukov. Demographic context in web search re-ranking. In X. wen Chen, G. Lebanon, H. Wang, and M. J. Zaki, editors, *CIKM 2012*, pages 2555–2558. ACM, 2012.
6. E. Kharitonov and P. Serdyukov. Gender-aware re-ranking. In W. R. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *SIGIR 2012*, pages 1081–1082. ACM, 2012.
7. P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley Longman, 2005.
8. J. Teevan, M. R. Morris, and S. Bush. Discovering and using groups to improve personalized search. In R. A. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, and B. B. Cambazoglu, editors, *WSDM 2009*, pages 15–24. ACM, 2009.
9. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. PAMI*, 32(9):1582–1596, 2010.
10. I. Weber and C. Castillo. The demographics of web search. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, editors, *SIGIR 2010*, pages 523–530. ACM, 2010.
11. I. Weber and A. Jaimes. Who uses web search for what: and how. In I. King, W. Nejdl, and H. Li, editors, *WSDM 2011*, pages 15–24. ACM, 2011.
12. D. Zeimpekis and E. Gallopoulos. TMG: A MATLAB toolbox for generating term-document matrices from text collections. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data:Recent Advances in Clustering*, pages 187–210. Springer, 2006.