Discovering user groups in professional image search

Theodora Tsikrika
Royal School of Library and Information Science, Copenhagen, Denmark
theodora.tsikrika@acm.org

Abstract

This study aims at gaining insights into user group identification in professional image search. The user groups are built by analysing the search logs recorded by a commercial picture portal for a sample of 170 users, in conjuction with the users' occupational and topical profile information, and a topical classification of the available images. Our analysis indicates that the examined groupings are meaningful and that there is variation among the groups in what people searched for and in what people considered relevant.

1 Introduction

Personalisation adapts information retrieval to the user context as this is delineated by the individual's explicitly or implicitly expressed interests. Data sparseness and ambiguity, particularly that of implicit indicators (e.g., queries and clicks), may lead though to the ineffective representation of user interests, especially for users with limited history. To this end, research has leveraged 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 in the same way for all members of the group [12].

One challenge in leveraging such group-based evidence is the identification of groups of related people. This study aims at developing an understanding of user group identification by investigating and comparing several features for building meaningful user groups and by examining the variation among the groups in what people searched for and in what they considered relevant. To this end, we analysed the search log data collected by the commercial picture portal of a European news agency for a sample of 170 of their registered users, in conjuction with the users' occupational and topical profile information and a topical classification of the available images. This news agency offers access to millions of photographic images to professional users, such as journalists, public relations agencies, editorial staff in newspapers and magazines, etc., all to be referred to as *journalists*.

2 Related Work

Several possible approaches to user segmentation have been explored, mainly in studies that ultimately aim to leverage group-based evidence for personalised or 'groupised' retrieval, or more generally context-aware re-ranking. The main features used for discovering user groups include occupational and topical evidence [12], demographic information [12, 7] with a particular focus on gender [8], geographical context [1], and reading level efficiency [2]. Their results show the usefulness of such evidence for improving retrieval.

Additional studies [14, 15] have investigated user segmentation outside the context of a particular retrieval application with the goal to gain a thorough understanding of how the search behaviour of specific groups

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

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, a finding that could be exploited in many applications, such as sponsored search.

Our study also focusses on analysing the searching behaviour of user groups, but in a context that differs to the search environments examined in all of the above work in at least one of the following aspects: (i) ours is a professional, rather than a web, environment, and (ii) it is oriented towards image, rather than text retrieval. Furthermore, we examine occupational and topical features for discovering user groups, similarly to [12], but consider the users' log activity rather than data collected through a user study. Finally, previous studies in journalistic search (e.g., [6, 4]) have mainly investigated the searching behaviour, the nature of queries, and the image selection criteria applied by individual users, rather than user groups.

3 Data Acquisition and Processing

Three data sources were used in this study: (i) a subset of the search log data collected by the commercial picture portal of a European news agency (`http://www.belga.be/`), (ii) the IPTC (*International Press Telecommunications Council*, `http://www.iptc.org`) classification of the images provided by the news agency, and (iii) profile information of a subset of their registered users; all data were made available to us under an NDA.

3.1 Search Logs Processing

The search log data used in this study were collected over a two year period (June 2007 – July 2009), with a three-month hiatus (October – December 2007). A sample consisting only of registered users logged into their account for whom profile information was available (see Section 3.3) was considered for analysis in this study. The logs recorded several search interactions, including users' query submissions and their clicks on selected images for further viewing and/or downloading (i.e., purchasing). Each log entry consists of a timestamp, the user's ID, and the submitted query. Click actions (viewing/downloading images) also logged the ID of the selected image.

Our sample was processed as follows in preparation for the analysis. 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]; session boundaries were identified when the period of inactivity between two successive actions exceeded a 30-minute timeout, similarly to [15]. Next, the submitted queries' text was 'lightly' normalised by converting it to lower case and removing punctuation, quotes, special characters, extraneous whitespace, URLs, and the names of major photo agencies. Also, 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 click. Our final sample thus contains 170 registered users who submitted a total of 198,410 queries (86,663 unique) and clicked on a total of 567,467 images (312,702 unique).

Table 1 lists some session statistics and their distribution across users. On average, our sample contains 547 sessions for each user with an average duration of about 17 minutes. The average number of queries/session is 4.4, close to the upper bound reported in previous web image search studies that employed the same session detection approach, where it ranges from 2.8 [13] to 4.8 [5] queries per session. However, it is slightly higher than what has been reported in professional image search [6] (3.3 queries per session); it should be noted though that no clear description of the session identification approach applied in that study is available. There are also 7.2 clicks per session on average by the users in our sample; no comparable statistics are available for similar image search studies (e.g., [6, 13, 5]). Finally, 71% of sessions resulted in at least one click, higher than what has been reported in web image search [13], where the same percentage is 56%. Overall, our analysis indicates that there are similarities with session characteristics reported in other analyses in journalistic and web image search.

Compared to an earlier analysis of a much larger sample of the same logs [4], where a 15-minute (rather than a 30-minute) timeout was employed, the session statistics in our sample when using a 15-minute timeout (3.6 queries/session, 6 clicks per session, 68% of sessions with at least one click) are comparable to those for the logged in users in the much larger sample (3.3, 5.5, and 62%, respectively); this indicates that the sample used in this study is representative of the user population of this commercial picture portal with respect to their session characteristics. Compared, though, to logs obtained from general-purpose web search engines, such as the much larger Yahoo! sample used in a similar study [15] that also detected sessions using a 30-minute timeout (where, on average, session duration is close to 7 minutes, with 2.4 queries/session, and 59% of sessions with at least

one click), our session statistics are different. Such differences have been observed before [5, 10], indicating that image search is potentially a more complex cognitive task than other types of search, and probably more so in the professional context investigated in this study. This suggests that potentially different features could be important for characteristing users', and thus groups', searching behaviour in different environments.

Table 1: Session statistics averaged across users.

	mean	stdev	10%-ile	90%-ile
# sessions	546.6	843.0	20.90	1415.6
duration (sec.)	1042.7	952.1	366.4	1974.0
queries/session	4.44	5.67	1.79	7.17
clicks/session	7.19	16.02	1.83	11.93
# 0-click sessions	129.2	188.1	5.90	392.3
% 0-click sessions	0.29	0.15	0.12	0.51
# 1-click sessions	100.6	164.1	2.00	290.7
% 1-click sessions	0.17	0.10	0.06	0.31

3.2 Image IPTC Classification

IPTC is a consortium of the world's major news agencies that provides news exchange formats to the news industry, and also creates and maintains concepts to be used as metadata to news objects; this allows for a consistent coding of news metadata. The news agency that provided us with the data uses, in addition to textual captions, the 17 IPTC *subject* codes (`http://cv.iptc.org/newscodes/subjectcode/`) listed in Table 2, i.e., the top level of IPTC's hierarchical newscodes, to describe the content of the images it provides.

Out of the 312,702 unique images clicked in our sample, 274,201 (87.8%) had been manually classified by the news agency's archivists. The performed classification is considered to have close to 100% precision. A manual inspection of some randomnly selected samples indicates that there is some noise, but its level appears to be low, though this cannot be accurately quantified. Some of this noise may be introduced by the requirement for strict classification to a single category and the inherent subjectivity in any such process.

Table 2 lists the distribution of the clicked images over the 17 IPTC subjects. Sports dominate, indicating a slight bias in the topical interests of the sampled users, as these are reflected by their searching behaviour. This is followed by politics and cultural topics 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.

Table 2: IPTC subject codes and the distribution of clicked images over them.

	IPTC subject codes		#images	% classified images	% all images
	Code	Name			
1.	ACE	Arts, Culture, & Entertainment	39,019	14.2%	12.5%
2.	CLJ	Crime, Law & Justice	10,461	3.8%	3.3%
3.	DIS	Disaster & Accident	6,862	2.5%	2.2%
4.	EBF	Economy, Business & Finance	17,128	6.2%	5.5%
5.	EDU	Education	648	0.2%	0.2%
6.	ENV	Environmental issue	2,075	0.8%	0.7%
7.	HTH	Health	2,148	0.8%	0.7%
8.	HUM	Human interest	13,489	4.9%	4.3%
9.	LAB	Labour	2,198	0.8%	0.7%
10.	LIF	Lifestyle & Leisure	2,702	1.0%	0.9%
11.	POL	Politics	42,375	15.5%	13.6%
12.	REL	Religion	2,369	0.9%	0.8%
13.	SCI	Science & Technology	1,911	0.7%	0.6%
14.	SOI	Social issue	1,497	0.5%	0.5%
15.	SPO	Sport	118,704	43.3%	38.0%
16.	WAR	Unrest, Conflicts, & War	8,242	3.0%	2.6%
17.	WEA	Weather	2,373	0.9%	0.8%
			274,201	100%	87.8%

3.3 User Profiles

The news agency's editorial staff provided us with the following information for each of their registered users: their affiliations (i.e., the name of the company they work for), the type of that affiliation (i.e., if it is a newspaper, a TV station, etc.), and some remarks in plain text regarding the topics of interest of that user (i.e., if they are

mainly interested in sports, politics, etc.). Table 3 lists the number of users affiliated with each company type, which shows that most journalists work for radio/TV stations, or magazines. Furthermore, based on the above information (affiliation, affiliation type, and remarks), users were manually classified to each of the three levels of the IPTC hierarchy (`http://show.newscodes.org/`), i.e., *subject* (level 1), *subject matter* (level 2), and *subject detail* (level 3), so as to reflect their topical interests; this classification is listed in Table 5 and is discussed next.

4 User Groups

This study groups users along two axes. The first relates to whether group membership is (i) determined *explicitly* by information provided by the users (to the news agency's staff), or (ii) inferred *implicitly* by their searching behaviour. The second relates to the *features* shared by group members: (i) *occupational*, relating to the jobs people have, and (ii) *topical*, relating to their interests. In particular, the investigated user groupings are: (i) two types of *explicit* groupings (*occupational* and *topical*), and (ii) three types of *implicit* groupings (all *topical*).

Explicit groups. *Occupational* groups consist of people with related jobs. In our case, it is assumed that people working in similar types of companies perform similar types of journalistic tasks. The occupational groups listed in Table 3 correspond to the **company type** grouping.

Topical groups consist of people who share an interest in a particular topic. Here, topical groups of users with shared interests are explicitly formed through the manual classification of users to each of the three levels of the IPTC hierarchy (see Section 3.3). The first grouping, denoted as **iptc user (level 1)**, is formed by considering only the top level IPTC classification, shown in the leftmost part of Table 5. Given the high percentage of users assigned to Economy, Business & Finance (EBF), a very broad category that appears to encompass a wide range of topics and therefore not being very discrimininative, a refined classification of the users belonging to EBF was applied and their second level IPTC classification was considered, shown in the middle part of Table 5. The second grouping, denoted as **iptc user (level 2)**, thus consists of 15 groups: 9 IPTC subject (level 1) groups, those listed in Table 5 excluding EBF, and the 6 IPTC subject matter (level 2) groups under EBF. Similarly, given the high percentage of users under the EBF/media class, a further refinement was applied to its members and their third level IPTC classification was considered, shown in the rightmost part of Table 5. The third grouping, denoted as **iptc user (level 3)**, consists of 19 groups: the 14 groups of the *iptc user (level 2)* grouping excluding the EBF/media group, and the 5 IPTC subject detail (level 3) groups under EBF/media. Some of the users in the EBF/media group had not been assigned to a third level class; these are all grouped under EBF/media/*_no detail_*. This third grouping achieves a less biased distribution of users across the groups.

Implicit groups. Topical groups of users with shared interests are implicitly formed based on the hypothesis that such users issue similar queries and/or click on similar images, e.g., with similar captions or IPTC codes.

The first grouping, denoted as **text–kmeans–*k***, is formed by applying *k-means* clustering on term vectors each corresponding to a user and representing their queries and the captions of their clicked images. To this end, the text of their queries and the captions of their clicked images are concatenated. The term vector is created after stemming, but without removing stopwords, and the term weights are estimated using a *tf.idf* scheme with normalisation. The k-means clustering uses the Euclidean distance and terminates when the objective function shows no further improvement; k-means ran 100 times for each k. Several different groupings were generated by varying the number of clusters k from 5 to 20. Both the term vector generation and the clustering were performed with the Text to Matrix Generator (TMG) Matlab toolbox [16].

The second grouping, **iptc clicks–kmeans–*k***, is formed by applying *k-means* clustering on vectors of the IPTC subject distribution of users' clicked images. Similarly to above, several different groupings were generated

Company type	# users	% users
agency	13	7.6%
government	5	2.9%
international organisation	2	1.2%
magazine	46	27.1%
newspaper	4	2.4%
private customer	20	11.8%
radio - tv	58	34.1%
website	22	12.9%
	170	100%

Table 3: Users affiliated with each company type.

IPTC subject	# users	% users
ACE	42	24.7%
CLJ	3	1.7%
DIS	1	0.6%
EBF	9	5.3%
HUM	12	7.1%
POL	46	27.1%
REL	1	0.6%
SPO	56	32.9%
	170	100%

Table 4: Groups based on the IPTC subject of the majority of clicked images.

Table 5: IPTC codes manually assigned to users to reflect their interests.

IPTC subject	#users	IPTC subject matter	#users	IPTC subject detail	#users
ACE	9				
CLJ	1				
EBF	117	EBF/business	10		
EDU	8	EBF/computing and IT	17		
HTH	2	EBF/construction and property	1		
HUM	12	EBF/economy	2		
LAB	2	EBF/finance	2		
LIF	2	EBF/media	85	EBF/media/*_no detail_*	39
POL	8			EBF/media/advertising	13
SPO	9			EBF/media/news agency	11
				EBF/media/public relation	1
				EBF/media/television industry	21
	170		117		85

by varying the number of clusters k between 5 and 20. The third grouping, **iptc clicks**, is formed by assigning to each user the IPTC subject code of the majority of their clicked images leading to the 8 groups listed in Table 4.

A comparison of the distribution of IPTC subjects manually assigned to users to reflect their interests (Table 5) with the distribution of IPTC subjects of their clicked images (Tables 2 and 4) shows a clear disparity. This is difficult to explain but may be due to a number of reasons. For instance, there might be a discrepancy between what users state when registering with the news agency based on their anticipation of what they will be working on and actual practice in their work life. Furthermore, journalists do not necessarily work on the same topic for a long period, and this effect might be more pronounced here given the long time period covered by our logs. Finally, journalists may select images to illustrate articles, reports, web sites, etc., which evoke associations rather than stress the (subject) content of such texts [9], and thus diverge from their topical interests.

Overall, four explicit and 33 implicit groupings are analysed. Topical user groups are the main focus as we aim to gain insights on the main feature employed by most personalisation approaches for discriminating among users: their topical interests. Further groups could be formed based on other features, including the sesssion characteristics presented in Section 3.1; this is left as future work.

5 Analysis

The above groupings are analysed to investigate how meaningful they are, i.e., whether group members are more similar to each other than to members of other groups, by examining the variation in what people searched for and in what people considered relevant. Further insights are gained by zooming in on particular clusters. First the method applied for evaluating our user segmentation outside the context of a specific application is presented.

5.1 Method

Evaluating these user groupings is akin to performing cluster validation; see [11] for a discussion on the notions presented next. Comparing individual groups within a given grouping or entire groupings can be performed in an *unspurevised* manner, by evaluating how well the groups fit the data without any reference to external information, or in a *supervised* manner, against known ground truth. Both these approaches are applied in our analysis and are briefly described next. Clustering(s)/cluster(s) are used interchangeably with grouping(s)/group(s).

Unsupervised cluster validation evaluates the goodness of a clustering based on inter– and intra–cluster pairwise proximity measures. In particular, the overall validity of a clustering can be expressed as the weighted sum of the validity of individual clusters, which is in turn measured either by inter–cluster *cohesion* expressing how closely related the objects in a cluster are, or by intra–cluster *separation* expressing how well-separated a cluster is from other clusters. Cohesion is defined as the average of the pairwise proximity values of all points within the cluster and separation as the average of the pairwise proximity values of each point within the cluster to all points in all other clusters. Average proximity in a clustering is computed based on all possible pairs.

Here, proximity is computed using the following similarity measures for user pairs: (i) the number of their *common queries*, normalised by the maximum such value observed in our sample (there are on average about 23 common queries per user pair in our sample), or (ii) the *Jensen-Shannon divergence* (a symmetrised KL divergence) between the IPTC subject distributions of users' clicked images. The latter is actually a distance metric and its value subtracted from 1 is employed instead[1]. Another similarity measure that could be used

[1]Jensen-Shannon divergence values range between 0 and 1 when the logarithm with base 2 is used, as done in this study.

is the number of common clicks in a user pair. However, our analysis showed that this would probably be an unreliable measure given our sample's very low numbers of shared clicks for users' common queries. This might be due to the highly dynamic and recency-oriented journalistic context, where image collections are constantly updated and users typically seek up-to-date information. Therefore, time-dependent relevance in conjunction with the long time period covered by our logs is a likely explanation for this phenomenon.

In addition to cohesion and separation, their combination, as this is reflected in the *silhouette coefficient*, is used. The silhouette coefficient of a clustering is defined as the average silhouette coefficient of all its points, while that of a point is computed using each of the similarity measures described above (see [11] for its definition).

Supervised cluster validation measures how well the constructed groups match a given ground truth. The following measures that evaluate the extent to which a cluster contains objects of a single class are used: (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. The entropy (purity) of a clustering is computed as the sum of the entropy (purity) values of each cluster weighted by its size.

5.2 Variation within Groups

Our analysis first explores how meaningful the 37 groupings are by examining whether the members in a group are more similar to each other than to members of other groups, using the similarity measures described above. Figure 1 (left, middle) clearly indicates that, in all cases, group members are more similar to each other than to users in other groups in terms of the queries they issue and the IPTC subjects of the images they click.

Starting with the explicit groupings, users affiliated with similar types of companies (*company type* grouping) appear to share more common queries, but to click on less similar IPTC subjects, compared to users manually classified as sharing common interests (*iptc user (level l)* grouping), at least for $l = \{2, 3\}$. These latter manual user classifications also appear to benefit by their refinements towards the deeper levels of the IPTC hierarchy with respect to their cohesion, but not their separation. Given that in our case only one IPTC subject (EBF) was refined, this indicates that indeed this top level subject is too broad to be discriminative, but that, on the other hand, its refined groups share topical interests not only among themselves, but also with users in other groups, most likely with those in the other groups obtained from the EBF refinement.

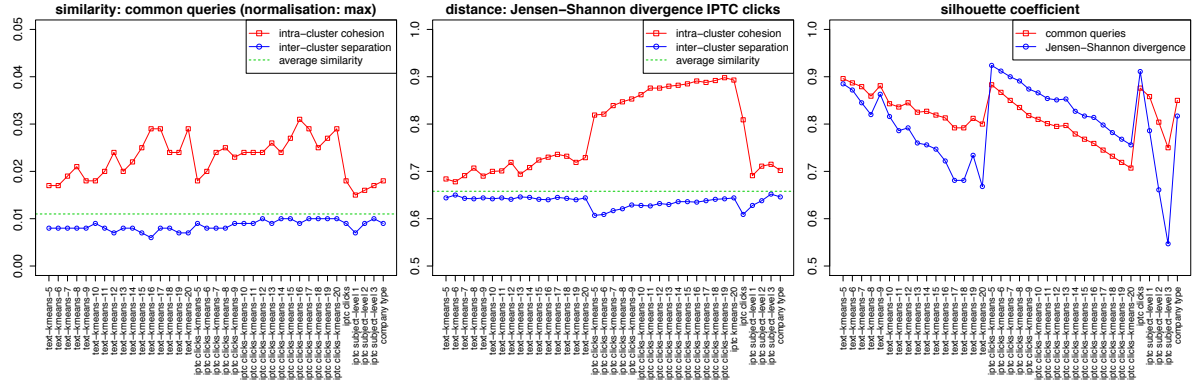

Figure 1: Variation in group membership for each of the 37 groupings: by comparing their intra–cluster cohesion and inter–cluster separation to the average similarity of all user pairs, using the *common queries* (left) and *Jensen-Shannon divergence* (middle) measures, and by their silhouette coefficient for the two measures (right).

For the implicit groupings, our analysis should take into account the relation that exists in some cases between the objective function used for the clustering and the similarity measure used for cluster validation. For instance, the objective function in the *iptc clicks–kmeans–k* clusterings and the *Jensen-Shannon divergence* measure for cluster validation are both based on the distribution of the IPTC subjects of the images clicked by users. This results in boosting these groupings' cohesion and separation values when this measure is applied, as shown in Figure 1 (middle). Similarly, but to a lesser extent, the objective function in the *text–kmeans–k* clusterings considers, together with the clicked images' captions, the users' queries, also considered by the *common queries* measure. Therefore, our focus is mainly on the validation of the *text–kmeans–k* clusterings using the *Jensen-Shannon divergence* measure, and the validation of the *iptc clicks–kmeans–k* using the *common queries* measure.

The results of this analysis in Figure 1 (left, middle) show that group members who clicked on images with similar IPTC subjects also issued more similar queries, and vice versa.

Figure 1 (right) plots the silhouette coefficient for all groupings, combining cohesion and separation. The effects of the relations between cluster validity measures and objective functions are also evident here. Generally, the most cohesive and well-separated groupings are those for lower k, and also the *iptc clicks* grouping.

Next, the entropy and the purity of each of the 32 implicit clusterings (*text–kmeans–k* and *iptc clicks–kmeans–k*) is examined in terms of users' classifications in each of the four explicit groupings and in *iptc clicks*. Regarding the explicit groupings, Figure 2 shows that all clusterings have the lowest entropy and highest purity for the *iptc user (level 1)* grouping, followed by the *iptc user (level 2)*, *company type*, and *iptc user (level 3)* groupings. This indicates that the clusters created in the implicit groupings mostly cluster together users who have been assigned the same IPTC subject, rather than working for the same company. This is to be expected though given the highly unbalanced data in the top IPTC level manual classification (see Table 5).

When using the *iptc clicks* classification, the entropy and purity of the *iptc clicks–kmeans–k* clusterings achieve their best values, as expected. In addition, though, also the *text–kmeans–k* clusterings have low entropy and relatively high purity with respect to *iptc clicks*. This indicates that groups consisting of users issuing similar queries and clicking on images with similar captions contain users that mostly select images with the same IPTC subject, thus further confirming the correlation observed above.

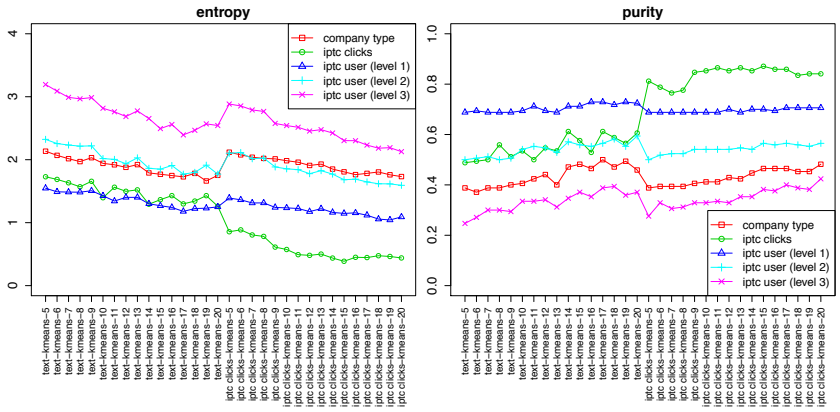

Figure 2: Entropy (left) and purity (right) of the *text–kmeans–k* and *iptc clicks–kmeans–k* ($k \in [5, 20]$) clusterings with respect to the user distributions in *company type*, *iptc user (level l)* ($l = \{1, 2, 3\}$), and *iptc clicks*.

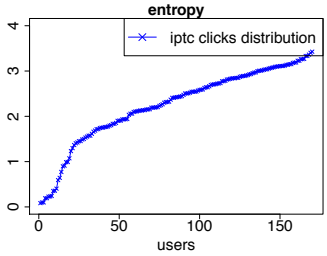

Figure 3: User entropy with respect to the IPTC subjects of their clicks.

5.3 Selected Clusters

To gain further insights into group membership, we selected one clustering to examine: *text–kmeans–10*, a clustering with relatively low entropy, high purity, and high silhouette coefficient. Table 6 lists for each cluster its distribution, entropy, and purity over the classes in *iptc clicks*, the two most frequent queries submitted by its members, and their average session statistics. For clusters with a dominant class and thus relatively low entropy (i.e., all clusters except 5 and 8), the most frequent queries are very representative of the dominant IPTC subject (with a Belgian focus, given their origin). For example, the most popular queries in clusters 1 and 9 are clearly relevant to culture/entertainment and sports, respectively, while those in cluster 2 relate to the *imperial and royal matters* category of IPTC's Human Interest subject. For clusters 5 and 8, where the distribution is equally split across a number of classes, an examination of their 20 most frequent queries shows that they cover several different subjects, indicating a more mixed membership. Regarding the session statistics, there is a clear outlier, while the rest are well below the sample's overall averages (see Table 1).

5.4 Individual Users

Finally, we have a closer look at individual users' searching behaviour and in particular at the distribution of the IPTC subjects of their clicked images. Figure 3 plots the entropy of that distribution for each user. Only about a fifth of our users have an entropy below 1 and about half have an entropy over 2. This indicates that many of

Table 6: For each cluster in *text–kmeans–10*, the table presents its distribution, entropy, and purity over the classes in *iptc clicks* (in bold the most frequent class), the two most frequent queries submitted by its members, and their average session statistics (duration, # queries/session, # clicks/session, and % zero-click sessions).

	ACE	CLJ	DIS	EBF	HUM	POL	REL	SPO	entropy	purity	queries	duration	#queries	#clicks	%zero-click
1	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	belgium royal arts / madonna	405.55	1.265	2.561	0.143
2	0.182	0.000	0.000	0.000	**0.727**	0.091	0.000	0.000	1.096	0.727	mathilde / prince laurent	57.55	0.243	0.208	0.046
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	1.000	boonen / anderlecht	317.65	0.644	0.935	0.011
4	**0.917**	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.414	0.917	madonna / michael jackson	75.87	0.295	0.701	0.018
5	0.000	*0.250*	0.000	0.000	*0.250*	*0.250*	*0.250*	0.000	2.000	0.250	maynard jackson / selys	242.20	1.279	1.132	0.191
6	0.000	0.000	0.000	**0.857**	0.000	0.143	0.000	0.000	0.592	0.857	bellens / reynders	80.80	0.475	0.215	0.036
7	0.300	0.025	0.000	0.025	0.037	**0.312**	0.000	0.300	2.010	0.312	leterme / obama	5.05	0.037	0.027	0.003
8	0.000	0.000	*0.500*	*0.500*	0.000	0.000	0.000	0.000	1.000	0.500	ghislenghien explosion / flood	2205.10	19.178	52.712	0.113
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	1.000	fellaini / kompany	335.40	1.679	2.266	0.111
10	0.121	0.000	0.000	0.000	0.000	**0.545**	0.000	0.333	1.374	0.545	obama / verhofstadt	27.90	0.118	0.146	0.008

the users in our sample click on images assigned to many different IPTC subjects; thus, their interests appear to cover several topics. Given the long period covered our searchlogs and the discussion in Section 4 on journalists' work practices, a more in-depth analysis that is time-based and also considers fuzzy clusterings is needed.

6 Conclusions

This work studied user group identification through the analysis of search log data collected by a commercial picture portal for a sample of 170 of their registered users, in conjuction with the users' occupational and topical profile information, and the topical IPTC classification of the available images. Overall, our analysis indicates that the examined groupings are meaningful, since groups members are more similar to each other than to users in other groups. Furthermore, it provides some support to the hypothesis that users who click on images with similar IPTC subjects also issue more similar queries, and vice versa, than the population at large. It also indicates that the relationship between group membership and issued queries and/or IPTC subjects of clicked images might be a good source of evidence for determining groups when these are not known a priori. This suggests that members of highly cohesive groups with respect to the above would probably benefit from a 'groupisation' approach. However, given the preliminary nature of this work, further investigations are needed for consolidating and generalising our findings. Finally, future work will follow a number of directions, including the use of the images' visual features for group identification, the consideration of semantic evidence for query similarity and classification, the exploitation of session information, and joint clustering using multiple features.

References

[1] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *Proc. of the 34th SIGIR*, 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 *Proc. of the 20th CIKM*, 2011.

[3] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12), 2009.

[4] V. Hollink, T. Tsikrika, and A. P. de Vries. Semantic search log analysis: A method and a study on professional image search. *JASIST*, 62(4), 2011.

[5] B. J. Jansen, A. Spink, and J. O. Pedersen. The effect of specialized multimedia collections on web searching. *J. of Web Engineering*, 3(3–4), 2004.

[6] C. Jörgensen and P. Jörgensen. Image querying by image professionals. *JASIST*, 56(12), 2005.

[7] E. Kharitonov and P. Serdyukov. Demographic context in web search re-ranking. In *Proc. of the 21st CIKM*, 2012.

[8] E. Kharitonov and P. Serdyukov. Gender-aware re-ranking. In *Proc. of the 35th SIGIR*, 2012.

[9] S. Ornager. The newspaper image database: empirical supported analysis of users' typology and word association clusters. In *Proc. of the 18th SIGIR*, 1995.

[10] S. Özmutlu, A. Spink, and H. C. Özmutlu. Multimedia web searching trends: 1997-2001. *Information Processesing and Management*, 39(4), 2003.

[11] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining, (First Edition)*, chapter 8. Cluster Analysis: Basic Concepts and Algorithms. Addison-Wesley Longman, 2005.

[12] J. Teevan, M. R. Morris, and S. Bush. Discovering and using groups to improve personalized search. In *Proc. of the 3nd ACM WSDM*, 2009.

[13] D. Tjondronegoro, A. Spink, and B. J. Jansen. A study and comparison of multimedia web searching: 1997–2006. *JASIST*, 60(9), 2009.

[14] I. Weber and C. Castillo. The demographics of web search. In *Proc. of the 33rd SIGIR*, 2010.

[15] I. Weber and A. Jaimes. Who uses web search for what: and how. In *Proc. of the 4th WSDM*, 2011.

[16] D. Zeimpekis and E. Gallopoulos. TMG: A Matlab toolbox for generating term-document matrices from text collections. In *Grouping Multidimensional Data*. Springer, 2006.