Mining query logs induced graphs

Alexandre P. Francisco[1*] Ricardo Baeza-Yates[2]
Arlindo L. Oliveira[1]

[1]INESC-ID / CSE Dept, IST, Tech Univ of Lisbon, Portugal
[2]Yahoo! Research Barcelona, Spain & Santiago, Chile

[*]Work done while visiting Yahoo! Research Barcelona.
Corresponding author: `aplf@ist.utl.pt`

Last modified: August 5, 2010

Abstract

The human interaction through the web generates both implicit and explicit knowledge. An example of an implicit contribution is searching, as people contribute with their knowledge by clicking on retrieved documents. When this information is available, an important and interesting challenge is to extract relations from query logs, and, in particular, semantic relations between queries and their terms. In this paper we present and discuss results on mining large query log induced graphs, namely click induced graphs, and how they contribute to query classification and to understand user intent and interest. Our approach consists on efficiently obtaining a hierarchical clustering for such graphs, which provide a hierarchical query folksonomy. We compare it against conventional taxonomies, namely the ODP categorization. Results obtained with real data show that the hierarchical clustering and the inferred query folksonomy provide interesting insights both on semantic relations among queries and on web users intent, being quite different from standard taxonomies.

1 Introduction

Nowadays the Web is the biggest representation of human knowledge, where people contribute with content either explicitly or implicitly. An example of an implicit contribution is searching, as people contribute with their knowledge by clicking on retrieved documents. Thus, queries submitted to search engines carry implicit knowledge and they can be seen as equivalent to tags associated to clicked documents. An interesting challenge is then to extract relevant semantic relations from query logs, which have several interesting applications. For instance, ranking algorithms, query recommendation systems and advertisement systems integrate such semantic information to improve their results.

In this paper we discuss query classification and meaning, and not URL tagging and folksonomies. Nevertheless, we use click-data to infer relationships and similarities among queries. Then, by finding closely related queries and relevant terms, we are able to define a hierarchical query folksonomy by associating tags to queries. As we should see, this approach may associate a tag to a query even if that tag does not make part of the query, leading to query contextualization. Our approach relies on graphs to represent relations both among queries and between queries and URLs. We start with the bipartite graph of queries and URLs, where a query q and an URL u are connected if a user clicked in the URL u that was an answer for the query q. We also know how many times a given URL u was clicked for each query and, thus, we weight each edge accordingly to click frequency. A second graph has queries as nodes and we add an edge between two queries whenever they share at least a common clicked URL. Each edge is also weighted by computing a similarity score between queries, such as a vector representation of the queries in the high dimensional space of all unique URLs. A more frequent approach is to define a similarity measure among queries ignoring the common clicked URLs. However it is more difficult to understand why queries are similar and it can add noise to data already noisy.

Within this line of work, graph mining techniques are crucial to uncover relations in query graphs. According to SearchEngineWatch.com, the number of queries of large search engines per day is of the order of hundreds of millions. By considering just a one day query log, the query graph would have tens of billion edges. Thus, analyzing such huge graphs is a hard task, that becomes even harder if we take into account similarity weights on the edges. On the other hand, the number of potential relations and their applications is huge.

Our study follows recent works on the analysis of query graphs [5, 17], which introduce the notion of click induced graph and present several results concerning semantic relations among queries. Here we propose three main contributions. First, given the existence of noisy relations among queries mostly caused by multi-topical URLs, we start by discussing how to detect such URLs, proposing a new heuristic. Second, we study how recent results on graph clustering can improve the extraction of semantic relations from query graphs and contribute to query classification. We tackle the problem of clustering click induced graphs, namely we discuss an efficient hierarchical clustering method for these large weighted graphs. We use a well known local optimization approach applied to seed sets, that may however fail if we choose the wrong seeds. Thus, we propose a suitable core enumeration procedure to select seed sets. Third, given a hierarchical clustering, we discuss the inferred semantic relations among queries and how the clustering can induce a query folksonomy. Note that although folksonomies are not usually hierarchical, in our case label specialization allows the creation of a hierarchical folksonomy.

To evaluate our approach we use a sample of a query log of a large search engine and we compare our results with a query classification obtained by mapping queries over the Open Directory Project (ODP) categories. The idea is to analyze how much of the knowledge expressed in queries is different from traditional topic classification.

2 Related Work

Most of the work on query similarity is related to query expansion or query clustering. Here we mention only the most closely related papers to our work.

Clustering similar queries is a common task in many applications such as query recommendation systems. Wen *et al* [24] proposed to cluster similar queries using four notions of query distance: (1) based on keywords or phrases of the query; (2) based on string matching of keywords; (3) based on common clicked URLs; and (4) based on the distance of the clicked documents in some predefined hierarchy. As the average number of words in queries is small and the number of clicks in the answer pages is also small [2], notions (1) and (2) generate distance matrices that are very sparse. For notion (4) we need a concept taxonomy and the clicked documents must be classified into that taxonomy as well, something that usually requires direct human intervention and that cannot be done in a large scale. Although notion (3) can generate also sparse distance matrices, the sparsity can be greatly reduced by using large query logs. Previous works have used notion (3), such as Befferman and Berger [7], or even variants combining (1) and (3) as well as other simpler features such as in Zaiane and Strilets [26].

Baeza-Yates *et al.* [3, 4] used the content of clicked Web pages to define a term-weight vector model for a query. They consider terms in the URLs clicked after a query. Each term is weighted according to the number of occurrences of the query and the number of clicks of the documents in which the term appears. Then, the similarity of two queries is equivalent to the similarity of their vector representations, using the cosine distance function. This notion of query similarity is based on common clicked URLs as (3) and has several advantages. First, it is simple and easy to compute. On the other hand, it makes it possible to relate queries that happen to be worded differently but stem from the same topic. Therefore, semantic relationships among queries are captured. More recently, Shen *et al.* [22] also used the notion (3) to cluster similar queries and build a query taxonomy. As Baeza-Yates *et al.*, they also consider the terms in the clicked documents instead of the terms in the queries. In this paper we represent each query in a high dimensional space, where each dimension corresponds to a unique URL, and the weights are defined accordingly to click frequency. This notion of similarity uses common clicked URLs and it was introduced by Baeza-Yates and Tiberi [5] to analyze a very large query log. They define semantic relations such as equivalence or specificity based on different set conditions among the set of clicked URLs. Using the ODP they found a precision up to 83% on the relations discovered and also that the ones not found were too specific to appear in ODP. More recently, Francisco *et al.* [17] have further studied the query graph generated by such similarity relations and they found that even a simple clustering approach can produce interesting results. In the present paper we further improve these results.

The work by Chuang *et al.* [10, 11, 12, 21] also uses query logs to build a query taxonomy to also cluster answers. However they do not use any user feedback, like user clicks. This idea of building a taxonomy based on queries is extended in [9], but this is not the same as building a taxonomy of the queries, which is what we call a query taxonomy or folksonomy. Later, Dupret and Mendoza [15] used the rank of clicked URLs to define relations among queries. They recommend better queries by generating query relations that can be associated to parts of a query taxonomy.

3 Click Induced Graph

Let $\mathcal{Q}$ be the set of queries and $\mathcal{U}$ be the set of URLs. Given a query $q \in \mathcal{Q}$, the cover of q is the set of URLs clicked by q. Let $\mu : \mathcal{Q} \to 2^{\mathcal{U}}$ be a function that maps each query q to its cover set $\mu(q) \subseteq \mathcal{U}$. The *click induced graph* $G = (V, E)$ is an undirected graph with queries as nodes and where exists an edge between two queries whenever they share at least one common clicked URL. Formally, $V = \mathcal{Q}$ and $E \subseteq V \times V$ is such that $(q_1, q_2) \in E$ if and only if $\mu(q_1) \cap \mu(q_2) \neq \emptyset$.

In what follows we will refer to the weighted click induced graph. Edges are weighted according to the cosine similarity of the queries they connect. Thus, for $(q_1, q_2) \in E$, the weight is given by

$$\sigma(q_1, q_2) = \frac{\sum_{u \in \mu(q_1) \cap \mu(q_2)} \rho(q_1, u)\rho(q_2, u)}{\sqrt{\sum_{u \in \mu(q_1)} \rho(q_1, u)^2}\sqrt{\sum_{u \in \mu(q_2)} \rho(q_2, u)^2}}, \tag{1}$$

where $\rho : \mathcal{Q} \times \mathcal{U} \to [0, 1]$ is a function such that $\rho(q, u)$ is the frequency ratio with which the URL u was clicked for the query q.

For experimental evaluation we considered a query log piece from a large search engine. The data was collected in April 2005 and contains 2,822,337 queries with at least one clicked URL and 4,927,980 different URLs. From these, only 660,910 URLs were clicked for more than one query and these are the relevant ones since we are interested in common clicked URLs. On average, each query has 2.39 distinct clicks and each URL is clicked by 1.37 distinct queries. Both click distributions, per query and per URL, follow a power law, with exponents 3.50 and 2.59, respectively. Queries comprise 554,380 different terms.

The main purpose of the click induced graph is to represent semantic relations between queries and to enable knowledge extraction. Semantic relations can however have low quality introducing noise. In what concerns the edge weights for the studied query log, we have that about 75% of edges are weighted with values bellow 0.5 and 50% with values bellow 0.273. Thus, there are many connections between queries which are not much similar. Most of this connections are due to URLs covering dubious topics, several topics or very general topics. These URLs are usually denoted as multi-topical, being examples many e-commerce and directory sites.

An approach to remove noise is to ignore contributions from multi-topical URLs. Baeza-Yates and Tiberi [5] suggested that multi-topical URLs are the ones that contribute more to edges with low weights. Then, we regenerate the click induced graph ignoring such URLs. Although this approach reduces the graph size removing the noise, we observed that URLs which contribute more to low weighted edges also may contribute more to high weighted edges. Moreover, we also observed a strong positive correlation between the number of queries covered by a URL and the number of contributions to edge weights. In Figure 1 we plot the geometric mean of the URLs weight contribution versus their size for our query log data. These results are due to the high number of queries for which a given multi-topical URL is the only clicked URL, generating many high weighted relations in the graph.

To solve this problem, we considered as documents the terms among the set of queries covered by each URL and we computed the tf-idf score for each term as usual. We observed that multi-topical URLs have a low average tf-idf score. This is true even when we select the high related queries for which those URLs were clicked. Therefore, we propose to compute the maximum tf-idf among the bag of terms associated to each URL and select the URLs with lowest score as multi-topical candidates. In Figure 1 we depict the maximum tf-idf score against URL coverage size for the query log analyzed. As we discuss ahead, this approach effectively reduces the size of the graph keeping its

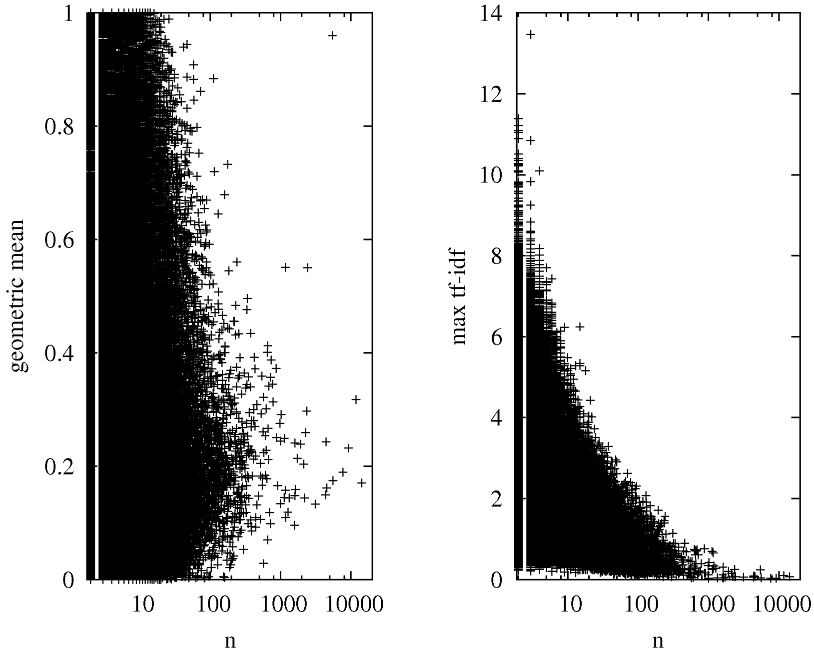

Figure 1: On left, URL weight contribution geometric mean versus URL coverage size n. On right, given the set of terms associated to the queries covered by each URL, URL maximum tf-idf score versus URL coverage size n.

properties, such as the size of the giant component and the weight distribution. We should note that this is consistent with previous results [5].

Next, we sorted the URLs by the maximum tf-idf score and we regenerate the click induced graph ignoring 0.05% of the URLs with lowest score, namely ignoring the 330 URLs with lowest tf-idf score. In Figure 1 we provide the distribution of tf-idf scores for the analyzed query log and, by selecting just 0.05% of the URLs we are filtering the click induced graph in a conservative way. Note that many of the selected URLs have a large coverage and, maybe unexpectedly, they are not spam URLs.

The resulting click induced graph has 23,177,430 edges, about 6.44% of the size of the full click induced graph. Since we continue having low weighted edges, we remove 10% of the edges with lowest score, all of them having a weight lower than 0.043. Thus, the filtered click induced graph has 20,974,257 edges and 1,648,649 connected components. The giant component contains 861,903 vertices and the second smallest connected component has only 64 vertices. There are now 1,474,249 singleton vertices. The degree distribution follows a power law with exponent 1.65. Therefore the approach to remove noise and multi-topical URLs dramatically reduces the size of the click induced graph, keeping its structure almost unchanged. This is an important fact since we can effectively reduce the noise without losing much information [5].

4 Hierarchical Clustering in Graphs

One of the hardest problems in graph mining is detecting graph community structure or graph clustering. The notion of community and the first formalizations of the concept have been proposed in the social sciences. Usually, communities are groups of vertices such that the number of edges within the groups is higher than the number of edges among different groups. The general aim of community finding and graph clustering methods is to detect meaningful divisions by inspecting the structure of the network. This problem has recently attracted a large interest and, for a deep review on this topic, we refer the reader to a review on complex networks by Boccaletti *et al.* [8] or a more recent survey on community finding by Fortunato [16].

In this paper we follow a two stage approach. We find a set of seed sets and, then, we apply a well known local optimization method. Several methods have been proposed based on the optimization of a given score [6, 19, 20, 23, 27], in particular to detect overlapping clusters based on global partition and local expansion [20, 23]. As pointed out by several authors, the main problem is how to choose seed sets. Usually, a well known spectral partitioning method is used to generate seed clusters, for instance multilevel bisection. Although the results are promising, such approaches inherit some of the drawbacks of standard multilevel methods when we are looking for overlapping communities. This problem becomes even harder when we have weighted graphs. To solve this problem, we propose a core enumeration method based on a vertex similarity score, where a core is a densely connected sub-graph which usually occurs within communities or clusters and that, by local optimization, leads to the full cluster.

4.1 Finding Cores

We could define the initial seed sets by taking many different approaches. Since we are interested in forming clusters of similar queries, a simple approach could be thresholding the edge weights. Another approach could be find the nearest neighbors. But we know that an URL may induce a clique in the graph and, in particular, it can induce a clique with high weights. If we follow the simple approach we could join two cliques even if they are connected by a single edge. Thus we take a different approach by trying to join similar queries, *i.e.*, we define cores based on vertex structure similarity. Let $G = (V, E)$ be a graph and $\sigma : E \longrightarrow \mathbb{R}_0^+$ the edge weight function. Given two connected vertices $(v_1, v_2) \in E$, their structural similarity takes values in $[0, 1]$ and is given by

$$\eta(v_1, v_2) = \text{avg}(v_1, v_2) \cos(v_1, v_2), \tag{2}$$

where $\text{avg}(v_1, v_2)$ is the weight mean among common neighbors, *i.e.*,

$$\text{avg}(v_1, v_2) = \frac{\sum_{w \in N(v_1) \cap N(v_2)} \sigma(v_1, w) + \sigma(v_2, w)}{|N(v_1) \cap N(v_2)|} \tag{3}$$

and $\cos(v_1, v_2)$ is a cosine similarity based score given by

$$\cos(v_1, v_2) = \frac{2\,\sigma(v_1, v_2) + \sum_{w \in N(v_1) \cap N(v_2)} \sigma(v_1, w)\sigma(v_2, w)}{\sqrt{1 + \sum_{w \in N(v_1)} \sigma(v_1, w)^2}\sqrt{1 + \sum_{w \in N(v_2)} \sigma(v_2, w)^2}}, \tag{4}$$

with $N(v)$ being the set of neighbors of v. The term $\cos(v_1, v_2)$ measures how similar are the two vertices with respect to common neighbors and respective weights. $\cos(v_1, v_2)$ takes value 1.0 whenever the vertices v_1 and v_2 share all neighbors, even if they are connected through edges with low weights. Thus, we introduced the term $\text{avg}(v_1, v_2)$ to distinguish common neighbors connected through low weighted edges from common neighbors connected through high weighted edges.

Given $\varepsilon > 0$, we say that $(v_1, v_2) \in E$ is a *core edge* if $\eta(v_1, v_2) \geq \varepsilon$. A set of vertices $C \subseteq V$ is a *core* in G if all $v \in C$ form a connected component composed only of core edges. By choosing different values for $\varepsilon > 0$, we can enumerate the set of cores in a graph at different resolutions. By considering the edges in decreasing order with respect to η, we obtain a hierarchy of cores. This method takes $O(|E| \max(\Delta, \log |V|))$ time, with Δ the maximum vertex degree: $O(|E|\Delta)$ time to compute the similarity for each edge; $O(|E| \log |V|)$ time to sort the edges in decreasing order; and $O(|E|)$ time to obtain the hierarchical clustering.

4.2 Local Optimization

Given a hierarchy of cores, we take each core as a seed set and we perform a local optimization step based on the local partition method proposed by Chung [13]. This method takes a seed set and uses a heat kernel [14] to expand it, which is a typical random walk and that was shown to provide good results [13]. Given a graph G, the transition probability matrix W of a typical random walk on G is defined as $W = D^{-1}A$, where A is the adjacency matrix of G (since G is undirected and weighted, A is symmetric and its entries are the edge weights), and D is a diagonal matrix with $D_{vv} = \sum_{w \in N(v)} \sigma(v, w)$. The heat kernel is then defined as $e^{-\alpha L}$, where $L = I - W$ with I the identity matrix. The parameter $\alpha > 0$ is known as the temperature and it plays an important role as the heat diffusion coefficient. We did several experiments and, within the scope of this paper, different values of alpha do not change much the results. Thus, in what follows, α is equal to 1.0. Given a preference vector p_0 obtained from a seed set, we use the following discrete approximation %beginequation$p_\alpha = p_0 \left(I - \frac{\alpha}{k} L\right)^k$, where k is the number of iterations. Yang *et al.* [25] used this approximation in a different context and they proposed a heuristic to find the minimum number of iterations for a given approximation error threshold. In particular, if the graph is connected, then p_α converges to the stationary distribution. However, we are not interested in this limiting distribution but rather in the distributions obtained after a small number of steps. Given a seed set, we define p_0 as the uniform distribution over the seed set and we simulate several heat kernel steps, computing the probability distributions p_α. After each step, we sort the vertices in descending order according to the degree-normalized probabilities $r_\alpha(v) = p_\alpha(v)/d(v)$, where $d(v) = \sum_{w \in N(v)} \sigma(v, w)$. This ordering defines a collection of sets $\{C_i\}_{i=1}^{\ell}$, where $C_i = \{v_j \mid 1 \leq j \leq i\}$ and ℓ is the number of vertices v such that $r(v) \neq 0$. We select the set C_i that minimizes the conductance

$$\Phi(C) = \frac{\partial(C)}{\min(\text{Vol}(C), \text{Vol}(V \setminus C))}, \tag{5}$$

where the volume $Vol(C)$ and the cut size $\partial(C)$ are given by

$$\text{Vol}(C) = \sum_{v \in C} \sum_{w \in N(v)} \sigma(v, w) \qquad \text{and} \qquad \partial(C) = \sum_{v \in C} \sum_{w \in N(v) \setminus C} \sigma(v, w), \tag{6}$$

respectively. Note that conductance is trivially minimized if C is V. Usually we are interested in a reasonable expansion of the seed set. In this paper we stop after finding the first local minimum. The cut sizes and the volumes for all sets C_i can be computed in $O(\mathrm{Vol}(C_k))$ time, by determining the change to C_i due to the addition of vertex v_{i+1}. This process is referred to as a *sweep* [1].

5 Induced Query Folksonomy

Given a hierarchical clustering on a graph, we show here how we obtain a folksonomy by associating to each node in the hierarchy the most relevant terms. Given a node in the hierarchical clustering tree, such association takes into account the queries in the leafs of the underlying subtree. We identify the most relevant terms for each internal node in the clustering tree as follows. We identify the level of the tree for the given node and we compute the set of queries associated to each node at that level by grouping together all queries in the subtree. Moreover, we identify the set of terms for each node simply by inspecting the set of queries. Those sets of terms become our documents and we infer the most relevant terms for each node by computing the tf-idf score for each term in each document. Those terms become the tags associated to the queries on that node.

The folksonomy labels are selected by sorting the terms by decreasing relevance. We can either select the first terms or just the ones scored above a given threshold. The later approach may conduce to unlabeled nodes, which we may prefer instead of bad quality labels. In particular, since click induced graphs are scale-free and have a giant component, internal nodes corresponding to the giant component, or even to part of it, have bad quality labels which do not bring relevant semantic information - see ahead.

6 Experimental Evaluation

We applied the hierarchical clustering method described in Section 4 to the filtered click induced graph. In Table 1 we provide several statistics for different snapshots of the hierarchical clustering. Since we removed the singleton vertices from the graph, we are considering 1,348,088 vertices. The degree distribution follows a degree power law and the graph contains a giant component, thus the giant cluster for $\varepsilon = 0$ was expected and it coincides with the giant component. In particular for $\varepsilon = 0$, the clusters are precisely the non-singleton connected components in the original graph. Moreover, we can see that the method effectively clusters the giant component. For instance, with $\varepsilon = 0.4$ the biggest cluster is much smaller, about 1.1% of the original giant component. Note also the low values for average conductance Φ, that core vertices are contained in non-singleton clusters and that the cores do not overlap initially.

6.1 Semantic Contextualization

Although the clustering is effective, we obtain many small clusters at each level. These correspond to loosely connected clusters that could appear connected if we consider larger query logs. Many are highly specific queries, such as “53545 clinic in janesville riverview wisconsin”, for which the search engine returns a low number of results and where the user clearly knows what he wants. Navigational queries also fall in this category, examples being “java.com www” or “slashdot.org”.

Table 1: Clustering statistics for different values of ε, where $\mathcal{C}$ is the set of non-singleton clusters, $\mathcal{S}$ is the set of singleton clusters, C denotes non-singleton clusters, $\hat{\Phi}$ is the clustering average conductance, n_{core} is the number of core vertices, $n_{\text{non-sing}}$ is the number of queries in non-singleton clusters and n_{overlap} is the number of queries in more than one cluster. The percentages refer to the increasing in the number of non-singleton queries and to the number of vertices in more than one cluster after optimization, respectively. The hierarchical clustering contains 1,348,088 distinct queries.

ε	$\lvert\mathcal{C}\rvert$	avg $\lvert C\rvert$	max $\lvert C\rvert$	$\lvert\mathcal{S}\rvert$	$\hat{\Phi}$	n_{core}	$n_{\text{non-sing}}$	%	n_{overlap}	%
0.0	174,400	7.73	861,903	0	0.00	1,348,088	1,348,088	0.00	0	0.00
0.1	213,557	6.73	768,174	27,524	0.04	1,225,791	1,320,564	7.73	114,723	8.68
0.2	228,550	6.38	629,416	81,621	0.07	1,104,657	1,266,467	14.65	181,833	14.36
0.3	224,683	5,58	245,050	213,357	0.09	962,193	1,134,731	17.93	112,549	9.92
0.4	210,532	4.94	9,421	346,550	0.08	815,791	1,001,538	22.77	34,108	3.41
0.5	180,812	4.84	1,385	496,583	0.08	656,720	851,505	29.66	21,020	2.47
0.6	147,228	4.50	1,331	696,712	0.07	507,507	651,376	28.35	9,342	1.43
0.7	103,553	4.27	1,174	909,602	0.05	353,592	438,486	24.00	2,958	0.67
0.8	82,701	3.62	235	1,049,433	0.03	254,514	298,655	17.34	619	0.21
0.9	61,792	2.92	113	1,167,521	0.01	168,320	180,567	7.28	30	0.01
1.0	48,547	2.29	14	1,237,095	0.00	110,993	110,993	0.00	0	0.00

Figure 2: One example of overlapping clusters in the hierarchical clustering after local optimization and in a snapshot for $\varepsilon = 0.5$.

From the overlaps we can infer relevant information about queries, namely about their ambiguity, context, topics and term polysemy. As an example, let us consider the overlap in Figure 2. We see that "nasm" appears in two different contexts, namely it is an acronym to both the National Air and Space Museum and the National Academy of Sports Medicine.

The obtained clusters provided also interesting insights with respect to web slang,

```
...
fights street video → iraq war pictures
fights street video → iraq helicopter civilian
fights street video → intelligence network homelandsecurity
fights street video → tape caught fights
...
free → virus airlines anti → slow computer virus
free → virus airlines anti → airlines airlines.com
free → virus airlines anti → virus anti free
...
tires goodyear bridgestone → firestone bridgestone tire
tires goodyear bridgestone → tires cooper dunlop
tires goodyear bridgestone → goodyear gemini car
tires goodyear bridgestone → goodyear tires tire
tires goodyear bridgestone → tires bridgestone truck
...
```

Figure 3: Examples of label specialization for clusters which provide also tag refinement for the query folksonomy.

namely term polysemy and semantic relations. By just looking at the cluster in Figure 2, we can infer that NASM has two possible meanings and that one of them should confer a kind of certification. By considering terms within clusters, we can detect that for example "windows", "mouse" and "wine" are polysemic terms. For instance, the term "wine" appears in several clusters together with terms such as "napa", "food", "magazine", and "noir" relating it to beverages. However, we find at least one cluster where it appears with terms such as "linux" and "windows", clearly relating it to the Wine translation layer for Unix like operating systems. An approach to identify term polysemy is to compare the bag of terms among overlapping clusters. Clearly, if a query is in two clusters but they share few terms, then the query shall be polysemic. Similarly, by analyzing similar words in the same cluster we can detect misspellings.

Given the hierarchical clustering described above, we build the induced query folksonomy. Note that the folksonomy we discuss here is entirely based on user interaction through a search engine. We do not add any other source of data in order to filter or improve it. As mentioned before, the click induced graph is scale-free and has a giant component. We computed the tf-idf score for all clusters at each level and the scores become meaningful only for $\varepsilon > 0.3$. For $\varepsilon \leq 0.3$, the tf-idf score for the giant component takes values between 0.05 and 0.07, and the most relevant term is "free". Thus, in our discussion we focus on the categories with more queries and at clusters for $\varepsilon > 0.3$, since they have higher tf-idf scores and are more informative.

The folksonomy is rather different from usual taxonomies, both because of the click induced graph structure and because of the type of categories found. Note also that in traditional taxonomies the topics are selected beforehand, while we do not have any topic pre-specified. We observed that category paths correspond most of the times to keywords meaningful for users, such as trademarks. It is interesting that, although we consider the undirected version of the click induced graph, we are able to detect query specialization through the hierarchical clustering (see Figure 3).

Nevertheless, some of the categories are somewhat strange. For instance, the second group of categories in Figure 3 joins anti-virus on computers with anti-virus on airlines. The term virus makes sense in both contexts, as it is usual to run such software on computers and biological virus are also a current trend within flights and travels. However, such categorization seems to be wrong and it may occur because of some URL badly clicked.

6.2 Comparison with ODP

Evaluating the query classification is difficult since it is very different from traditional directories. In this section we try to compare it with the ODP[1] in order to understand how different are these two ways of expressing knowledge. We mapped all queries over the ODP categories, obtaining several category paths for each query. Then we compare the ODP paths with the induced folksonomy.

Since we have millions of queries, we had to perform such mapping offline. We download the ODP data and we fed it to the Lucene[2] search engine. The ODP data consists of a large set of URL entries, each one with a category, a title and a description. The ODP data set we considered for this paper contains 4,595,111 URL entries and 763,535 distinct categories. We took each URL entry as a document and we indexed all fields, *i.e.*, URL, category, title and description snippet. Lucene was configured to search over all fields and, for each query, to return several categories ranked by relevance. We used the default Lucene scoring function, which combines the Vector Space Model and the Boolean Model to determine the relevance of documents [18]. Note that we do not obtain categories to all queries. By inspecting Table 2, we see that 297,818 queries, 10.55%, are not mapped. If we compare with the click induced graph, we have that 67% of these queries are singleton queries. Moreover, 68% of the queries with a score lower than 1.0 as reported by Lucene are also singleton queries in the click induced graph. This is consistent with our observations about the singleton queries, that many of them are ambiguous and uninformative.

The folksonomy labels are not comparable to the categories in the ODP mapping since they are not topic based. Thus, we evaluate the clusters by comparing the common ODP path prefix among the queries. Given two queries q_1 and q_2, we select the two most similar ODP category paths p_1 and p_2, i.e., the ones which share the longest common prefix $\pi(p_1, p_2)$. Then we compute the score

$$\sigma_{\mathrm{odp}}(p_1, p_2) = \frac{|\pi(p_1, p_2)|}{\max\{|p_1|, |p_2|\}}, \tag{7}$$

where $|\cdot|$ denotes the path length. The ODP score for a given cluster is the average of the score σ_{odp} for all pairs of queries in that cluster.

For all snapshots of the hierarchical clustering at different values of ε, more than 50% of the clusters have an ODP score higher than 0.5. Since we do not obtain ODP categories for all queries, many clusters have an ODP score of 0.0. In our experiments, depending

[1] http://www.dmoz.org/

[2] http://lucene.apache.org/

Table 2: Query distribution over the ODP top categories. In this table we map each of the 2,822,337 queries to a single category, the category with highest score.

Category	Queries	%	Category	Queries	%
Adult	105,123	3.72	News	7,672	0.27
Arts	367,702	13.03	Recreation	99,224	3.52
Business	141,512	5.01	Reference	56,112	1.99
Computers	157,313	5.57	Regional	605,483	21.45
Games	69,830	2.47	Science	83,100	2.94
Health	74,261	2.63	Shopping	86,758	3.07
Home	70,152	2.49	Society	187,106	6.63
Kids and Teens	50,539	1.79	Sports	80,522	2.85
			World	282,110	10.00

on the value of ε, we have 16% to 30% of clusters with an ODP score equal to 0.0. We also have 30% to 39% of the clusters with an ODP score equal to 1.0. It is interesting that these clusters are small in the number of queries and that they appear independently of ε. Thus, we may infer that either these clusters are well defined or they are meaningless. It is interesting that this fact supports our previous observation that these clusters usually consist of either navigational queries or ambiguous and uninformative queries.

In order to analyze the remaining clusters, we select the 1000 biggest clusters at different depths, *i.e.*, for different values of ε. Note that, for $\varepsilon \leq 0.9$, the number of clusters with an ODP score of either 0.0 or 1.0 among the 1000 selected is 0%. Figure 4 depicts the ODP score values for the 1000 biggest clusters at different snapshot levels. Note also that the first snapshots have a higher average score, because for $\varepsilon \leq 0.3$ exists a giant component and the remaining clusters are rather small. As we mentioned before, the giant component vanishes for $\varepsilon > 0.3$. Thus, after we cluster the giant component, the score increases with the hierarchical clustering depth, revealing that clusters at higher depths have better quality. This is also supported by the tf-idf scores.

7 Final Remarks

Queries submitted to search engines can be viewed as an expression of the knowledge of the users. In this paper we propose methods to better understand the users intent and interest. The main goal is the efficient analysis of large query logs and, in particular, of click induced graphs. First we discussed how to filter out noise caused, for instance, by multi-topical URLs. In particular we proposed a method to detect such URLs based on the analysis of the queries for which URLs were clicked. Second we devised and applied a hierarchical clustering method for weighted graphs. This method was shown to be effective and the results revealed effective semantic relations between queries. Namely, by building an induced folksonomy, we were able to identify query contextualization and specialization. Another interesting result is the fact that our approach, based only on click-through data generated by the users, provides a query classification much different from the one expected by traditional directories. This points out how hard is query

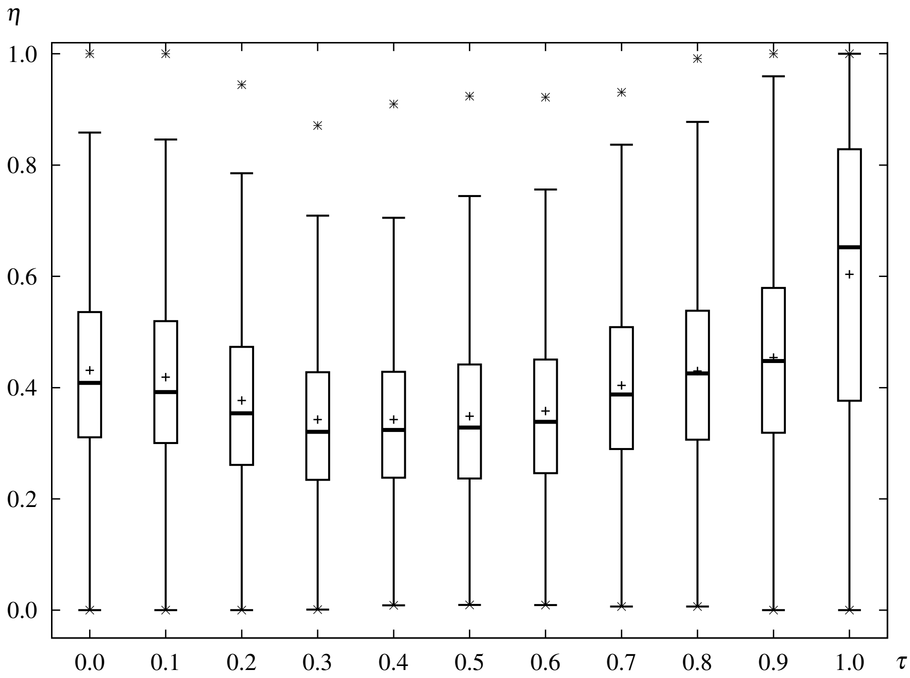

Figure 4: ODP score statistics for the 1000 biggest clusters at different snapshot levels, *i.e.*, different values of ε. For each snapshot, the box plot details the σ_{odp} minimum ($*$), the lower, median, and upper quartile, the maximum ($*$) and the mean ($+$).

classification and how highly relevant is the implicit knowledge found in query logs.

The quality of the results can be improved by incorporating more data, *i.e.*, by using larger logs, since more data will consolidate the relations obtained. The efficiency of the proposed methods makes them applicable to much larger graphs, thus making them suitable for the analysis of larger logs and for the extraction of semantic relations from less frequent queries. We should note that, for the log piece analyzed, all tests were run in common laptop. Since the methods are easily parallelizable, we are able to analyze much larger logs on high-end systems.

Keeping applications on sight, it is important to note that both the edges and the queries have probabilities associated and provided by the hierarchical clustering algorithm. Thus, we have confidence measures that are crucial to rank the relations among queries and their cluster membership.

References

[1] Reid Andersen and Kevin J. Lang. Communities from seed sets. In *World Wide Web (WWW'06)*, pages 223–232. ACM, 2006.

[2] R. Baeza-Yates. Applications of web query mining. In *European Conference on Information Retrieval (ECIR'05)*, volume 3408 of *LNCS*, pages 7–22. Springer, 2005.

[3] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query clustering for boosting web page ranking. In *Advances in Web Intelligence (AWIC'04)*, volume 3034 of *LNCS*, pages 164–175. Springer, 2004.

[4] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in a search engine. In *EDBT Workshops*, volume 3268 of *LNCS*, pages 588–596. Springer, 2004.

[5] R. A. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *SIGKDD*, pages 76–85. ACM, 2007.

[6] J. Baumes, M. Goldberg, and M. Magdon-Ismail. Efficient identification of overlapping communities. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 27–36, 2005.

[7] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *SIGKDD*. ACM, 1999.

[8] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[9] P.-J. Cheng, C.-H. Tsai, C.-M. Hung, and L.-F. Chien. Query taxonomy generation for web search (poster). In *CIKM*, 2006.

[10] S.-L. Chuang and L.-F. Chien. Towards automatic generation of query taxonomy: A hierarchical query clustering approach. In *IEEE International Conference on Data Mining*. IEEE, 2002.

[11] S.-L. Chuang and L.-F. Chien. Automatic query taxonomy generation for information retrieval applications. *Online Information Review*, 27(5), 2003.

[12] S.-L. Chuang and L.-F. Chien. Enriching web taxonomies through subject categorization of query terms from search engine logs. *Decision Support System*, 30(1), 2003.

[13] F. Chung. The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735, 2007.

[14] Fan R. K. Chung and S.-T. Yau. Coverings, heat kernels and spanning trees. *Electr. J. Comb.*, 6, 1999.

[15] G. Dupret and M. Mendoza. Automatic query recommendation using click-through data. In *IFIP World Computer Congress (WCC'06)*. Springer, 2006.

[16] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010. `arXiv.org:0906.0612`.

[17] A. P. Francisco, R. Baeza-Yates, and A. L. Oliveira. Clique analysis of query log graphs. In *SPIRE'08*, volume 5280 of *LNCS*, pages 188–199. Springer, 2008.

[18] O. Gospodnetic, E. Hatcher, and M. McCandless. *Lucene in Action*. Manning Publications, second edition, 2009.

[19] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.*, 11, 2009.

[20] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-define clusters. `arXiv:0810.1355`, 2008.

[21] H.-T. Pu, S.-L. Chuang, and C. Yang. Subject categorization of query terms for exploring web users' search interests. *JASIST*, 53(8), 2002.

[22] D. Shen, M. Qin, W. Chen, Q. Yang, and Z. Chen. Mining Web Query Hierarchies from Clickthrough Data. In *AAAI'07*, pages 341–346. AAAI Press, 2007.

[23] F. Wei, W. Qian, C. Wang, and A. Zhou. Detecting Overlapping Community Structures in Networks. *World Wide Web*, 12(2):235–261, 2009.

[24] J. Wen, J. Mie, and H. Zhang. Clustering user queries of a search engine. In *Proc. of the 10th International World Wide Web Conference*. W3C, 2001.

[25] Haixuan Yang, Irwin King, and Michael R. Lyu. Diffusionrank: a possible penicillin for web spamming. In *SIGIR'07*, pages 431–438. ACM, 2007.

[26] O. R. Zaiane and A. Strilets. Finding similar queries to satisfy searches based on query traces. In *Efficient Web-Based Information Systems (EWIS)*, 2002.

[27] S. Zhang, R. S. Wang, and X. S. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1):483–490, 2007.