Exploiting the Wisdom of the Crowds for Characterizing and Connecting Heterogeneous Resources

Ricardo Kawase, Patrick Siehndel, Bernardo Pereira Nunes,
Eelco Herder and Wolfgang Nejdl
Leibniz University of Hannover & L3S Research Center
Appelstrasse 9, 30167 Hannover, Germany
{kawase, siehndel, nunes, herder, nejdl}@L3S.de

ABSTRACT

Heterogeneous content is an inherent problem for cross-system search, recommendation and personalization. In this paper we investigate differences in topic coverage and the impact of topicstopics in different kinds of Web services. We use entity extraction and categorization to create 'fingerprints' that allow for meaningful comparison. As a basis taxonomy, we use the 23 main categories of Wikipedia Category Graph, which has been assembled over the years by the wisdom of the crowds. Following a proof of concept of our approach, we analyze differences in topic coverage and topic impact. The results show many differences between Web services like Twitter, Flickr and Delicious, which reflect users' behavior and the usage of each system. The paper concludes with a user study that demonstrates the benefits of fingerprints over traditional textual methods for recommendations of heterogeneous resources.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous—*Classification, Navigation*

Keywords

Fingerprints; Classification; Comparison; Domain independent; Wikipedia

1. INTRODUCTION

When searching on the Web for a particular topic, many different kinds of resources can be found, varying from tweets or news items on this topic to movies that have this topic as a keyword. Which resources are found, depends on whether one uses a general-purpose search engine or a specific site such as Twitter[1] or IMDb[2]. As an example, a query on 'Farming' at Twitter may lead to tweets as 'Five reasons why urban farming is the most important movement of our time', while IMDb suggests the 1957 BBC series 'Farming' and even the movie 'There Will Be Blood'.

User interests are as diverse as the topics covered in Web content. However, it is unlikely that a user who is interested in concrete topics such as sustainable farming methods also likes to watch fictional movies that happen to be situated in the countryside. This raises several issues for cross-system search, recommendation and personalization. For instance, besides regular Web content, Google[3] usually includes images and videos in the search results, which in many cases seems to be a shot in the dark. Still, it is likely that more general user interests, for example in sports or in culture, will be reflected in preferences for news items as well as books or movies. Conversely, some topics may be more represented in one ecosystem than in the other. For example, one would expect that 'Politics' is less prominent in Flickr[4] pictures than in Twitter messages.

In this paper, we investigate the differences in *topic coverage* in different kinds of Web sites. Furthermore, we investigate the *impact* of a topic on user appreciation: are movies on agriculture more or less popular than people who tweet on this topic? Our approach relies on the assumption that heterogeneous resources will have similar scores on the coverage of more general categories, such as agriculture, health and politics. Thus, we propose a method to generate *fingerprints* for objects that allow for meaningful comparisons between heterogeneous domains.

The most important characteristic of our fingerprint approach is that it has a limited, yet broad coverage of topics, based on Wikipedia[5] top categories that are maintained by the overall agreement of millions of contributors. Fingerprints provide users a sense making categorization that is digestible and manageable. While other approaches like clustering and LDA provide means for categorization and recommendation of items, they do not support the end user in understanding or configuring parameters.

Our proposed fingerprint is composed of a 23-sized vector that corresponds to the 23 main top categories of Wikipedia. Thus, for each category we assign a weight that represents its relevance for the given object. After all categories have been weighted, the fingerprint is created as a histogram, which characterizes a given object. Due to the collaborative nature of Wikipedia and the large use as a source of knowledge by the Web users, we have adopted its categories as our

[1] http://twitter.com/
[2] http://www.imdb.com/

HT'14, September 1–4, 2014, Santiago, Chile.

http://dx.doi.org/10.1145/2631775.2631797 .

[3] http://www.google.com/
[4] http://www.flickr.com/
[5] http://www.wikipedia.org/

knowledge base. Wikipedia currently contains over 4 million articles that have manually been categorized: one article may belong to one or more categories, and categories to one or more parent categories. However, although we use Wikipedia, any other classification scheme could be used along with our method, for instance, ODP[6], OpenCyc[7] or YAGO [19].

The process for creating the fingerprints is divided into 3-step process chain: (a) entity extraction; (b) categorization and (c) profile aggregation. Briefly, for any given object, our technique first recognizes its entities. After that, the entity categories are extracted and finally aggregated (following a weighting rule), creating the objects' fingerprint.

We validate and demonstrate our proposed approach with a set of experiments and analyses: (i) We validate our fingerprinting approach by comparing it with manual categorization (Section 4), (ii) we expose differences in topic coverage between several online systems (Section 5), (iii) we show the influence of topics on resources' *impact* (Section 6), and finally, (iv) we perform a user study that shows how fingerprints improve recommendations for heterogeneous resources (Section 7).

The results show that the 23 main categories of Wikipedia provide a solid base for reasoning about differences in topic coverage between, for example, Twitter and IMDb. Furthermore, the analysis of topic impact shows interesting differences in user appreciation of resources related to topics such as politics, nature and law, as expressed by the number of followers in Twitter and movie ratings in IMDb. As shown by the results of the user study, this opens the way for automatically identifying the most promising sites for finding resources that are relevant, but not necessarily directly related, to a topic and to incorporate this in the search results.

2. RELATED WORK

Ontologies and categories are commonly used as domain models in the field of user modeling and recommender systems [4]. Corresponding user models are represented as overlays of the domain model, in which the values represent the user's knowledge of or interest in a concept. Knowledge or interest levels are usually estimated based on user actions, such as the content of visited pages or the keywords of user queries. Propagation techniques, such as spreading activation [18], are used to ensure that evidence of interest in a particular concept also affects related concepts, such as its parents or children.

Wikipedia is a popular knowledge base for classification, categorization and even recommendations. Wikipedia is constantly refined by contributors and each article is assigned to a number of categories which are hierarchically organized, creating an implicit ontology [21]. In fact, numerous previous works leverage the use of Wikipedia categories. For example, Köhncke and Balke [11] exploit Wikipedia categories in order to generate useful descriptions for chemical documents. In their work, they identify chemical entities in documents and extract the categories of these entities. The resulting categories, combined with a tailored ontology, provided chemical documents with a better description (tag-clouds) than terms from a domain-specific ontology. The main difference from our work is that they deal with a very specific domain, and, they are not interested in co-relating objects with other fields. Thus, they only use the direct categories assigned to the entities, not exploring the category graph.

Chen et al. [5] exploit Wikipedia categories to improve Web video categorization. Their approach relies on identifying Wikipedia concepts in videos and exploiting the associated concepts' categories. Their outcomes describe a small improvement in the categorization task, however, in their work, Wikipedia concepts are manually identified from titles and tags of videos, and they manually performed a syntactical analysis to discriminate classes of concepts.

Blognoon [7] is a semantic blog search engine that leverages topic exploration and navigation. Blognoon provides faceted navigation for blog posts based on Wikipedia concepts. The main idea of the authors is to relate blog posts by the means of common concepts and, to some extent improve the exploration and serendipity in the blogosphere. Their work aligns with our goals of generating implicit relation between objects (in their case blog posts) by exploiting Wikipedia concepts. Unfortunately, the authors do not expose any evaluation. They claim that their query suggestion method, which is based on Wikipedia popularity, is more effective than alphabetical order. However, Wikipedia popularity of articles has been proven to be an ineffective foundation for recommendation [9]. Moreover, we believe that usability could be improved if categories were used instead of plain concepts.

Our approach of reducing the user profile to 23 topics differs significantly from the work of Michelson and Macskassy [14]. In their work, they propose a similar approach to annotate tweets with Wikipedia articles; but instead of considering all parent categories, they traverse the category graph only '5 levels deep'; they assume that a five stage traversal is sufficient to reach categories that are general enough for a user's profile. The limitation of their assumption is that a user's classification may have an unlimited number of categories, thereby preventing profiles from having a normalized length and comparison among all items.

Abel et al. [1] presented similar strategies to enhance Twitter user profiles, however their topic-based profile is built upon topics related to different types of news events. In our work, we consider the topics (categories) of each detected Wikipedia entity, thus the categories describe a wider area of fields. Moreover, they use as knowledge base the OpenCalais[8] ontology, which is a document categorization system that mainly focuses on news events.

Finally, regarding topic graph walk strategies (see Section 3), the method proposed by Kittur and Chi [10] to relate articles to categories is very similar to our approach. The main difference is that our approach is more focused on applications and not limited to articles inside Wikis. While the authors relate one article to the top-level category with the shortest path (or more if there is more than one shortest path) to see which content is inside Wikipedia, our approach relates articles to several top-level categories. This allows us a better comparison of profiles, due to the increased number of weights for each top-level category - in our case, no topic information is discarded.

[6] Open Directory Project - `http://www.dmoz.org/`

[7] `http://www.opencyc.org/`

[8] `http://www.opencalais.com`

3. FINGERPRINTS

As explained in the introduction, we aim to compare heterogeneous objects, based on fingerprint profiles of these objects. A common approach is to create a vector space model in which each field contains a score on a particular term or category.

There are several ways for selecting the terms or categories to be used for the vector. An IR approach would be to select the most frequent terms, excluding stop words. However, as our aim is to compare heterogeneous objects, it makes more sense to use an existing and well-accepted ontology or categorization.

There are many good candidate ontologies or knowledge bases, including YAGO, WordNet[9], and SUMO[10]. We decided to use the well established Wikipedia corpus as a semantic knowledge base. Wikipedia is arguably the most accessed reference Web site and each of the more than 4 million existing articles are manually classified by human curators to one or more categories. Additionally, categories are organized in a graph in which sub-categories reference to top-level categories. The English Wikipedia has a total of 23 top-level categories (*Main topic classifications*), which we use to represent a profile[11].

The creation of semantically enhanced profiles consists of three stages. During the first stage, *extraction*, entities are extracted from a given textual object. We first annotate the object to detect any mention of entities that can be linked to Wikipedia articles. For this purpose, we use the WikipediaMiner[15] service as an annotation tool. First, detected words are disambiguated using machine learning algorithms that take the context of the word into account. This step is followed by the detection of links to Wikipedia articles. Only those words that are relevant for the whole document are linked to articles. The goal of the whole process is to annotate a given document in the same way as a human would link a Wikipedia article.

In the second stage, *categorization*, we extract the categories of each entity that has been identified in the previous step. For each category, we follow the path of all parent categories, up to the root category. In some cases, this procedure results in the assignment of several top-level categories to a single entity. Following the parent categories (which are closer the root category), we compute values of distance and siblings categories, resulting in each entity receiving 23 categories' scores. In fact, there are different approaches that can be applied to walk Wikipedia's category graph. To achieve best results and accurately assign weights to each of the 23 categories, we experimented different graph walk and weighting strategies. A detailed evaluation is provided in Section 3.1.

Finally, in the *aggregation* stage, we perform a linear aggregation over all of the scores for a given object in order to generate the final profile.

3.1 Category Computation

We used the Wikipedia category graph for relating one article to the 23 main Wikipedia categories. The dataset we used contains 593,125 different categories. Each of these categories is linked to one or more of the main categories. Table 1 shows some statistics of the used graph.

[9] http://wordnet.princeton.edu/

[10] http://www.ontologyportal.org/

[11] http://en.wikipedia.org/wiki/Category:Main_topic_classifications

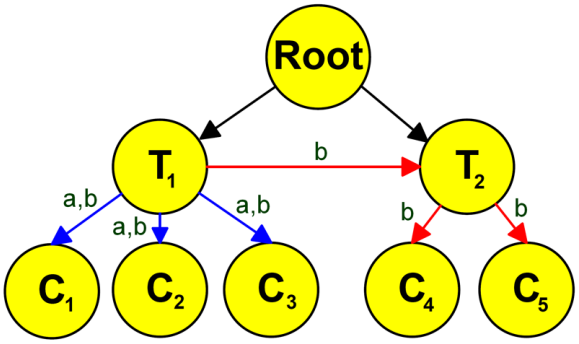

Figure 1: Subcategories of TC_1 for strategies 'A' and 'B'.

We used two different graph walking algorithms for computing the relation of a category to the main categories. Both strategies follow a top-down approach that pre-computes main category weights for each article. The main difference between the two approaches is the size of the generated tree for each main category. The relation of an article to the main categories is based on a depth-first walk through the Wikipedia category graph: the algorithm remembers the distance from the root node, and follows only sub-category links of which the distance is larger (*strategy A*) or equal (*strategy B*) to the current distance to the root node.

Figure 1 shows a small graph that consists of a root, two top-level categories (T_i) and 5 normal categories (C_i). When strategy 'A' is applied on this graph, category T_1 will contain all articles that are related to the categories C_1, C_2, C_3 and T_1. The category T_2 will not be part of C_1 because there exists another way with equal length from the root to T_2. When strategy 'B' is used on this graph, all categories will be seen as part of T_1.

By following only links that match this pattern, we make sure not to include the entire category graph (and all articles) for each main category. Additionally, we avoid loops by storing visited nodes and not visiting these nodes again. For the subcategories that are reachable through the category graph, we get the corresponding articles that belong to the categories. With this approach, we get a relation map in which every category is related to many articles and, in which most articles are related to many categories.

A basic profile (fingerprint) for an object consists of weights for all of the main categories. The final weight θ of a topic $t \in T$ (top 23 Wikipedia categories) for an object $o \in O$ is given by Equation 1:

Table 1: Statistics on the Wikipedia Category Graph.

# of Categories	593,125
# of Category-Subcategory links	1,306,838
avg. # of Subcategories	2.2
# of Page-Category Links	11,220,967
avg. # of Pages per Category	18.9

$$\theta(o_i, t_k) = \sum_{e \in o_i}^{l=|e|} \left(\sum_{c_j(e_l)}^{j=|c_j(e_l)|} w(c_j, t_k) \right), \quad (1)$$

where e are the entities annotated in a given object o, $c(e)$ are the Wikipedia categories for $e \in o_i$ and w is a weight given to the link between a category c_j and a top-category t_k. In Section 4, we define the weight used in our experiments (see Equation 2).

3.2 Resource Fingerprints

With the proposed approach, we are able to generate fingerprints for different sources of information and in different domains. This subsection explains how our approach can be applied to generate fingerprints for user-generated content, and for tags that are associated with a particular resource or user. In addition, we explore the benefits of generating fingerprints for movies (Subsection 3.2.3).

3.2.1 User-content Fingerprint

In the last years, social networking has become the most prominent online activity: the most popular social networks, including Facebook[12], Twitter, Myspace[13], aggregate over a billion users. As a result, research interest in the area of social networks has grown considerably. User modeling, link prediction, sentiment analysis, community analysis, sociology and many other areas of Web Science are examples of research fields that exploit the public (and private) data available from such networks.

User-content fingerprints are based on the contents produced by a user (in our experiments, all tweets posted by a user). In order to generate a user fingerprint we utilize the content posted by the user as the input corpus for the *Extraction*, *Categorization* and *Aggregation* steps. The resulting fingerprint represents the users main interests in terms of the 23 Wikipedia top categories. As a downside, the fingerprints are not detailed enough to be useful for generating recommendations or other kinds of personalization. However, this 23-size vector has the advantage that the profile is human understandable and allows for easy comparison between users.

3.2.2 Tag-Based Fingerprints

Another application area is the generation of fingerprints for tagged resources. Considering the associated tags of a given resource or user, we use them as input for the profiling process. In this case, instead of exploiting the whole text of a resource that inevitably introduces noise, the resulting fingerprint is based solely on entities identified by the tags.

Tags are mainly applied for describing the content of an item in order to facilitate the organization and management of the resources, and for making search and retrieval more effective [8, 20]. Additionally, tags enhance the visibility of community content by associating related items with the same annotation(s) [3]. Thus, fingerprinting items based on tags allows our approach to be applied to any folksonomy, even on those where resources do not have a textual representation (e.g. images and videos), which results in profiles that are more concise and less noisy.

[12] http://www.facebook.com/

[13] http://www.myspace.com/

3.2.3 Text-Based Fingerprints

To exemplify text-based fingerprints, let us consider movie descriptions. Although a movie is often classified by its genre, rarely there is a content-based classification of it. In the field of movie recommendation, most approaches make use of features that are based on co-occurrence of movies with actors and genres, user ratings, contextual and temporal information, together with collaborative filtering [12]. Recommenders that exploit the actor-movie-genre network are able to provide movie recommendations that hold similar characteristics. Collaborative filtering - based on user data and ratings - is able to provide very good recommendations, but in many cases for dissimilar movies (e.g. people who liked "The Manchurian Candidate", a political thriller, may also like "A Beautiful Mind", the dramatization of a mathematician biography, even though the movies have nothing in common).

By creating fingerprints of movies, we generate sense making profiles that are solely based on the content, i.e. the movie description. Our example movie "The Manchurian Candidate" would, based on its description, have 'Politics' as its main category. The benefits of such profiles are twofold. First, the reduced representation of topics of interest is based on a well-established knowledge base that effectively aggregates the wisdom of the crowds. In this way, fingerprints are comparable among any different entity type. Second, the profiles are human comprehensible, thus, any person is able interpret a fingerprint and understand the rationale behind it.

4. PROOF OF CONCEPT

In this section, we describe a experiment to evaluate the quality of the profiling methods through the recognition of entities in a text.

4.1 Experimental setup

In order to validate the applicability of the profiling method, we use articles from the Wikipedia corpus itself. In Wikipedia, articles are manually annotated with categories. The idea is to utilize these categories as the input for our method (in this case, starting from stage two, *Categorization*) and use the output as ground truth. As a result of this profiling method, we have a 23-sized vector, representing the fingerprint of a given article, which are solely based on the article's own categories.

The validation comes with a comparison of these profiles against the ones generated by applying the whole method. Therefore, the evaluation will measure the similarity of profiles generated by the existing categories of an article against the profiles generated by the categories of articles mentioned in an article.

Given the fact that manually assigned categories are descriptive, we aim to demonstrate that it is possible to categorize textual objects through the extraction of mentioned entities. This experiment is divided in two different stages. In the first stage, we aim to show that our approach leads to a good description of the main category of an article. Therefore, we selected articles that have already been annotated with one of the main categories. This set of 1444 different Wikipedia articles is then processed based on their categories, to relate it automatically to the top categories. This experiment will show that the categories of articles men-

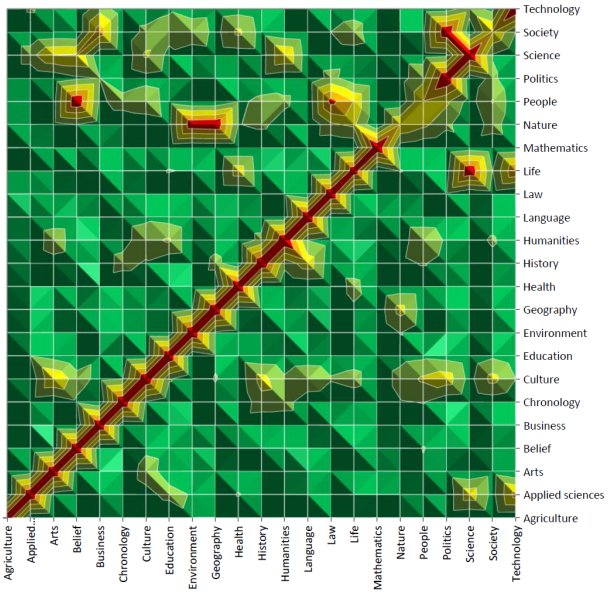

Figure 2: Relation between automatically assigned top category and ground truth.

tioned inside *containing articles* can be used to define the categories of the *containing article*.

In the second stage, we select a set of random articles from Wikipedia. With the second experiment we aim to evaluate if the categorization process leads to similar results when used with just the text of an article, or the categories directly related to an article. The set of articles contained in the set of the first article is very high in the category graph. Therefore, most of the articles in this set have very strong relations to just one of the main categories. The second set is randomly chosen and contains also articles which have strong relations to more than one of main categories. Additionally, we run the experiments using both graph walking strategies described in Section 3.1, in order to select the best performing one.

4.2 Experimental Results

The diagram in Figure 2 shows to which category articles from the different top categories are related. The diagonal line indicates that most articles from a certain top category are also classified as part of this category. Additionally, we see that there are strong relations between some of the categories. For instance, articles of the category 'Nature' are often classified with a high value in the categories 'Environment' and 'Geography'. The map was generated by using strategy 'B' together with a weighting scheme. The weighting is based on the distance of the article's categories to the root category and the probability of an article belonging to a certain top category.

There are big variances between the different categories. Categories like 'Mathematics', 'Agriculture' or 'Chronology' are relatively weakly represented. This leads to a classification in which these categories are underrepresented as well.

To achieve a more precise classification, we calculate the weight of the top categories taking into account the relative probability of an article belonging to one of the main categories. Additionally, we assume that a longer distance to one of the main categories can be interpreted as a weaker relation to that category. The calculation is shown as Equation 2,

$$w(t_k, c_j) = \frac{1}{P(t_k)} * \frac{1}{\delta(t_k, c_j)} \quad (2)$$

where $P(t_k)$ indicates the popularity of a given top-category and δ is the distance of a category c_j to the top-category t_k. To measure the performance in this experiment, we calculate the average rank of the correct main category inside the profile vector. For strategy 'A' we achieve an average rank of 2.9863 and for strategy 'B' we achieve 3.502. For the success@k we got very similar values for both strategies as shown in Table 2. To analyze how the performance changes when taking articles which are not so high in the Wikipedia category, we performed the same experiment with articles that belong to categories one level below the top categories. Overall, this dataset contained 33,262 different articles. The results are shown in Table 3 and as we can see, the overall performances of both algorithms are still good.

Table 2: Results of generated article category relations for articles of main categories

	Strategy A	Strategy B
Avg. Ranking	2.9863	3.5024
Success @ 1	0.5044	0.5073
Success @ 2	0.6568	0.6456
Success @ 5	0.8367	0.815
Success @ 10	0.9531	0.9085

Table 3: Results of generated article category relations for articles of subcategories of main categories

	Strategy A	Strategy B
Avg. Ranking	3.5198	4.0619
Success @ 1	0.4458	0.4102
Success @ 2	0.586	0.5585
Success @ 5	0.791	0.7627
Success @ 10	0.9314	0.8955

Beside the analysis of articles which are close to main categories, we performed an experiment to measure how the approach works for random articles. In order to select articles that contain enough content, we only selected articles with at least 20 inlinks, 30 outlinks and a minimum text length of 1000 characters. Overall we selected 10,000 different articles and applied our categorization method once on the content of the article, and once on the categories directly related to article. Since there are not necessarily any main categories directly related to the article, we measured the performance by means of cosine similarity between the generated profile based on the content and the profile based on the categories. The results of this experiment are shown in Table 4. As strategy 'B' performed better for random articles, we used this strategy for the remaining experiments in this paper.

5. TOPIC COVERAGE

As explained in the introduction, it is likely that there are differences in topic focus between Web services. In this section we use the fingerprint approach for identifying such differences in four distinct services: Flickr, Delicious[14], Twitter and IMDb, making use of an extensive dataset. As will

[14] http://delicious.com/

Table 4: Similarity results between the generated categories and the ground truth for 10000 random articles.

Depth of which Article	Strategy A cos-sim	Strategy B cos-sim	Number of Articles
ALL	0.8603	0.9275	10000
0-3	0.7314	0.8641	91
4	0.84	0.8897	536
5	0.8744	0.9214	2962
6	0.867	0.9320	2885
7	0.8631	0.9346	1904
8	0.8444	0.9390	1015
9	0.8208	0.9418	485
10	0.8081	0.9397	92
>10	0.7246	0.8689	30

Table 5: Datasets Statistics

	Data	Users	Identified Articles
Twitter	86,244 tags	1,574	32,569
Flickr	12,271,742 tags	14,450	5,341,331
Delicious	890,062 tags	2,005	558,409
IMDb	275,784 descriptions	-	1,351,433

be discussed in more detail in the remainder of this section, overall coverage per topic is quite consistent between all systems. However, when looking at the relative coverage of each topic, it becomes clear that, for instance, 'Mathematics' has a relatively high coverage in Delicious and Twitter, but - as one would expect - is less well represented in Flickr and IMDb.

5.1 Datasets

The tag-based datasets that are used in our evaluation were collected by Abel et al.[2] for the Mypes user profiling service. Flickr and Delicious data consist of tags that were assigned to resources, respectively pictures and bookmarks. In Twitter, the data consists of hashtags used in Tweets. As explained in subsection 3.2.2, these tags are the input to identify Wikipedia articles about the entities related to the resources, which on their turn were used for generating the fingerprints. For IMDb the input for generating the fingerprints was the descriptions of movies. Table 5 shows some statistics about the used datasets.

5.2 Fingerprints in different domains

We applied our profiling method to all users in the datasets described in the previous subsection. The collection of all users discriminated by each of the systems gives us an overall fingerprint for each domain (movies in the case of IMDb). All four systems receive a relative similar fingerprint profile. All have a very broad coverage in the categories 'Society' , 'Life' and 'Culture'. To make the results of the different systems easier to compare, we normalized the values based on the category with the highest coverage (in all cases 'Society').

For a better understanding of the coverage difference in each domain, we calculated the variation of topic coverage of all systems per category. Figure 3 quantifies the difference between each topic in each system and the global mean. The results show interesting aspects that uncover users' behavior

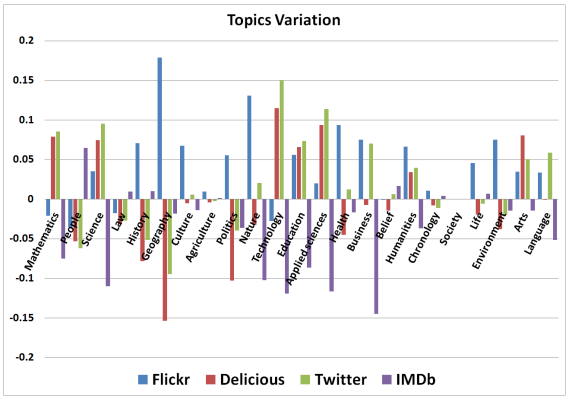

Figure 3: Variations in topic coverage in different systems.

and the usage of each system. For instance, Flickr is used to publish and tag pictures, many of these are tagged with locations, leading to a peak at the 'Geography' category. Other topics like 'Technology' are less covered in Flickr, but show a higher coverage for Twitter and Delicious. We also see that the graphs for Delicious and Twitter are very similar for most of the categories while the graphs for IMDb and Flickr show higher differences.

Many more differences can be observed in Figure 3, but a detailed analysis is beyond the scope of this paper. In general, the tendencies reflect expected differences between the different services.

5.2.1 *Topic Breakdown*

The fingerprinting approach is not limited to the generic top-level categories, but can also be used for breaking down a topic. To illustrate this, we also analyzed what the coverage of deeper categories looks like. Figure 4 shows how the different systems cover the subcategories of the 'Culture' (the plots do not include all subcategories of 'Culture', only those with significant coverage or variation). A closer look at the variations of each system against the global average shows that, for instance, Twitter and IMDb have strong relations to 'Entertainment' while Flickr shows peaks at 'Cultural spheres of influence', 'Cultural history' and 'Political culture'. The highest peak in 'Cultural spheres of influence' can be explained by the fact that this category has many subcategories which are related to geography. The peak in 'Cultural history' can be explained with tags on pictures of landmarks, a very representative set in Flickr [6].

6. TOPICS IMPACT

Given the differences in focus between domains and systems, it is likely that user appreciation of a tweet or a movie on a topic will be different. In other words, movies on certain topics may have a significantly lower or higher average rating in IMDb. Similarly, users who tweet on certain topics may have more or less followers than others.

6.1 Datasets

We applied our profiling method to the IMDb dataset from October 2012, consisting of over 2.3 million items (movies, tv series, etc). In total, we generated fingerprint profiles for

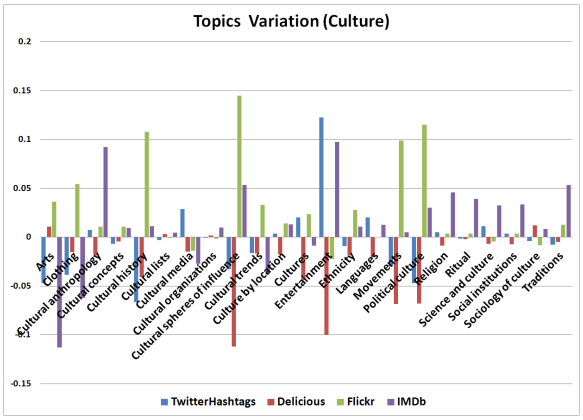

Figure 4: Variations in topic coverage of 'Culture' subcategories in different systems.

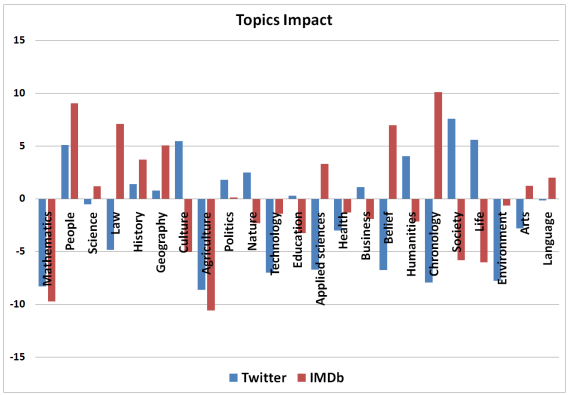

Figure 5: Percentual impact factors of topics on number of followers on Twitter and movies' ratings on IMDb.

275,784 movies - we ignored all identified episodes descriptions of series and other items apart from movies. With the movie fingerprints, we are now able to suggest movies that deal the same topics, without the compulsory attachment of genre, actors or ratings, and that are still interesting for the user. In the movies scenario, we use ratings as a success indicator.

While for movies higher ratings arguably indicate *better* movies, in the Twitter-user scenario, we used the number of followers as a parameter of *success* (the higher the number of followers the 'bette r' the user). Obviously, there are exceptions: celebrities on Twitter get millions of followers without posting anything interesting. Lim and Datta [13] propose an approach that involves identifying celebrities that are representative for a given topic of interest. In their work, they define a celebrity as a user that has more than 10,000 followers. Given this premise, we computed the impact factor of Wikipedia topics in Twitter based on the profiles of 1776 users - from these users we had information on the number of followers (average 74.7) and there were no celebrities. The resulting impact factors for Twitter users are depicted together with IMDb impact factors in Figure 5.

6.1.1 Topic-Based Movie Ratings

Arguably, ratings are the most prominent feature for recommending movies. As we have both ratings and topic weights for every movie, we can analyze the influence of topics in movies ratings.

To check the topic influence on ratings, we compute the difference between, on the one hand, the average percentages of topic distribution multiplied by movies' ratings, and, on the other hand, the average percentages of topic distribution multiplied by 6.54 (the global average of all ratings in IMDb).

To illustrate the analysis, imagine that there are 100 *Technology* movies and 100 *Politics* movies. The average distribution considering only these two categories is 50% for each. Now, let us assume that all *Technology* movies have ratings of 7.5 stars while *Politics* have ratings with 4.5 stars. By multiplying the distributions by the ratings, ($100 \times 7.5 = 750$ and $100 \times 4.5 = 450$), and the average rating ($100 \times 6.54 \cong 650$), and calculating the differences ($750 - 650 = 100$ and $450 - 650 = -200$), the absolute variation of 300 indicates a 33.3% positive impact factor of *Technology* and a 66.6% negative impact factor of *Politics*. We calculated the topic impact factor based on the ratings of all movies that had at least 1000 votes - in total 16,374 movies.

The topic impact factor ω is then given by Equation 3:

$$\omega(t_k) = \sum_{i=0}^{|m|}(\theta(m_i, t_k) \cdot \gamma(m_i)) - \sum_{i=0}^{|m|}(\theta(m_i, t_k) \cdot \frac{\sum_{i=0}^{|m|} \cdot \gamma(m_i)}{|m|}), \quad (3)$$

where the function θ gives the weight of topic for a movie m (see Section 3) and γ is the IMDb rating of a movie m.

As depicted in Figure 5, the topics 'Mathematics', 'Agriculture', 'Culture', 'Society' and 'Life' have a stronger negative impact on the movies ratings - or, in other perspective, movies that deal with those subjects are usually rated lower than others. These differences in topic impact largely match the variations in topic coverage, as discussed in the previous section. It should be noted that the differences in average movie rating may not be directly related to the topic per se: it may well be the case that the average movie on, for instance, 'Agriculture' is produced with a smaller budget and targets a particular audience - this in contrast to popular movie topics such as 'Law' or 'History'.

6.2 Topic Impact in Different Domains

To draw a comparison between people's interests and movie topics, we calculate the same *topic impact factors* for Twitter users. Not surprisingly, Twitter users tend to follow people who talk about popular topics as 'People', 'Society', 'Life' and 'Culture'.

Figure 5 also shows some interesting contrasts between both domains. People like movies about 'Law', but usually do not follow people who tweet about this topic - people tend not to be fond of lawyers. Additionally, 'Chronology', 'Applied Sciences' and 'Belief' seem to be topics that produce enjoyable movies, but are great turn-offs in Twitter.

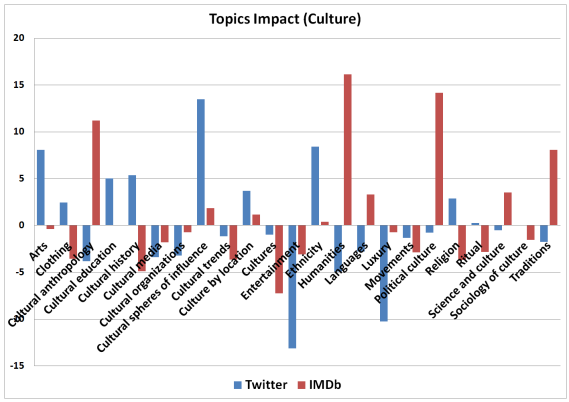

Figure 6: Impact factors (in percentage) of *'Culture'* subcategories on number of followers on Twitter and movies' ratings on IMDb.

6.2.1 *Topic breakdown*

Similar to the topic breakdown in Section 5, we apply impact analysis on subcategories of the main category 'Culture'. Figure 6 shows the impact factors for these subcategories - we only display those that have an influence higher than 1%. The chart shows interesting differences between both domains, where 'Cultural Anthropology', 'Humanities', 'Political Culture' and 'Tradition' have a positive impact for movies while in Twitter 'Arts', 'Cultural Spheres of Influence' and 'Ethnicity' provides the most positive impacts. Without going into further detail, the differences clearly show that users tend to rate movies on certain topics differently than tweets on the corresponding topics.

7. USER STUDY

This section presents the evaluation process used to validate our approach in terms of cross-domain recommendations. For this, we perform a user evaluation using a crowdsourcing platform to collect feedback. The goal is to compare recommendations given by the fingerprint approach against a text-based approach. The idea behind this study is to validate the usefulness of the fingerprint profiles to recommend heterogeneous resources in comparison to traditional text-based approaches. Specifically, the setup of our user study is to recommend movies that are relevant, but not necessarily directly related, to a given textual resource.

7.1 Datasets

In order to collect a useful dataset, we used the OAI-PMH protocol[15] to harvest resources that contain informative or educational content. We focus on repositories that provided an OAI-PMH target, among others *12Manage*[16], *INSEAD*[17], *LSE Research Online*[18].

After harvesting these different open repositories, we selected a random set of documents that were written in English and that contained at least 500 characters in its description. In total, we collected 1,416 resources to undergo our fingerprint method. For the movies dataset, we used the same as described in Section 6.1, however we selected only movies that are annotated with the genre 'documentary' (31,991 in total), which are assumed to provide interesting facts on a given topic, to be rather informative and in many cases entertaining.

7.2 Approach and Baseline

In order to generate recommendations, we used cosine similarity between the fingerprints. Thus, given a learning object and its fingerprint, we rank the movies according to their fingerprint' cosine similarity. As a result, for each resource, a ranked list of 'contextualized' movies is produced. With the purpose of comparison, we also generated rankings based solely on textual similarities.

To measure the textual similarity among the resources and movies, in our study, we used *MoreLikeThis*, a standard function provided by the Lucene search engine library[19]. *MoreLikeThis* calculates similarity of two documents by computing the number of overlapping words and giving them different weights based on TF-IDF [16]. *MoreLikeThis* runs over the fields we specified as relevant for the comparison - in our case the description of the resource and the movies' plots - and generates a term vector for each analyzed item (excluding stop-words).

To measure the similarity between items, the method only considered words that are longer than 2 characters and that appear at least 2 times in the source document. Furthermore, words that occur in less than 2 different documents are not taken into account for the calculation. For calculating relevant items, the method used the 15 most representative words, based on their TD-IDF values, and generated a query with these words. The ranking of the resulting items is based on Lucene's scoring function which is based on the Boolean model of Information Retrieval and the Vector Space Model of Information Retrieval [17].

7.3 User Task

We set up our evaluation on CrowdFlower[20], a crowdsourcing platform. With CrowdFlower, we are able to reach a broad and unbiased audience to judge our outcomes. The task posted for the participants consisted of evaluating the relevance and relatedness between a resource and a movie. Each participant was presented with the description of the resource and the description of the top-ranked recommended movies (with the same descriptions as used for the fingerprinting process). After reading the descriptions, participants were asked the following two questions:

- *Q1: Do you think that the suggested movie is relevant for the given document?*
- *Q2: In which degree the movie is related to the main topic of the document?*

The responses were registered using a 5-point Likert scale model. The first question aims at measuring the quality of the movie recommendations in terms of informational value. The second one aims at uncovering the actual topic-based relatedness of a movie and a resource. These answers are

[15] http://www.openarchives.org/pmh
[16] http://12manage.com/
[17] http://knowledge.insead.edu/
[18] http://eprints.lse.ac.uk/
[19] http://lucene.apache.org/core/old_versioned_docs/versions/3_4_0/api/all/org/apache/lucene/search/similar/MoreLikeThis.html
[20] https://www.crowdflower.com/

not necessarily dependent: a movie may not be relevant, but still topic-wise related. For example, a document on the economical crisis in Greece is topically related to the movie 'My Big Fat Greek Wedding', but they are arguably hardly relevant for each other.

7.4 Results

In total, we had 60 participants in our evaluation. These participants evaluated 606 pairs of movie recommendations. The responses were evenly distributed between fingerprints and the text-based approach (303 judgments for each). In general, for the fingerprint-based strategy, 74% of the participants *agreed* or *strongly agreed* on the *relevance* of the recommendations. In contrast, the positive agreement results for the relevance of the text-based strategy sums up to only 55% (see Table 6). Regarding relatedness, the results turned out to be quite similar. Both strategies produced around 44% related (>3) recommendations.

To extend our analysis, we calculated the Pearson's coefficient of correlation between the first and the second question, resulting in 0.52 for the fingerprints strategy and 0.80 for the text-based. In both cases, we see a high correlation, specially for the text-based approach. The main reason is that the text-based approach is unable to capture different aspects other than explicit terms in the description. Thus, if it produces a relevant result, most probably it will also be related. On the other hand, fingerprints identify relevance without relatedness. In fact, results show that for the fingerprint approach, in 13.9% of the judged pairs, the participants stated that the movies were relevant (*agree* or *strong agree*) but not related (relatedness 1 or 2). For the opposite case, where movies were related (relatedness 4 or 5) but not relevant (*disagree* or *strong disagree*), it only happened in 1.3% of the judgments. Respectively, the numbers for the text-based approach are 7.2% and 1.3%. These numbers suggest that even though a movie is unrelated to the main topic of a document, it might still be relevant.

To summarize, the results show that the fingerprint approach produces significant ($p<0.05$) better recommendations in terms of relevance. Our *fingerprinting* approach is able to identify the context of a document (using the 23 main topic categories of Wikipedia) on a higher level of abstraction. In contrast, a text-based approach is not able to identify these general topics and relies solely on term-to-term identification. In general, text-based approaches fail to identify latent topics in rather short descriptions. Fingerprints overcome this problem by efficiently recognizing relevant and contextualized entities in the objects' descriptions.

8. CONCLUSION

In this paper, we presented a fingerprint-based approach for comparing different kinds of resources in different domains. Fingerprint-based profiles are created by extracting entities from free text, categorizing them into one of Wikipedia main categories and aggregating the results into one profile. We validated this approach by comparing manually assigned main categories of Wikipedia articles with automatically found categories. Fingerprint-based profiles of user content, textual descriptions and tags were used for identifying differences in topic coverage and topic impact in different domains and systems, among which Twitter, Flickr, IMDb and Delicious.

Experimental results show that the Fingerprint-based approach is able to quantify and visualize differences in focus of these systems, such as the focus of Twitter messages on recent events, with entertainment as a main interest. We also showed that certain topics receive significantly higher or lower ratings in a system. As an example, movies about agriculture usually receive lower ratings in IMDb and people who tweet about agriculture have a less-than-average number of followers. As we discussed, these tendencies need not to be caused by the topic per se.

There are numerous applications where fingerprints can be applied: assisting systems and users to disambiguate queries, to control diversity in results and to overcome language differences. Fingerprint-based profiles are especially useful in situations when apples need to be compared to oranges, a situation that is not uncommon. As an example, user profiles, which are used for recommendation and personalization, are usually specific to the domain and the system, and therefore cannot easily be applied elsewhere. Fingerprint-based profiles are less precise than regular user profiles, but provide a good basis for creating an initial interest profile in cold-start situations.

Fingerprints also provide users insight in differences in focus for different systems. For systems such as IMDb, knowledge on significant differences in user ratings per topic can be used for compensating for these differences, in order to better cater queries for specific (niche) topics or for users with a high interest in these specific topics.

To illustrate our approach, we deployed an online system[21] that allow users to generate fingerprints for Twitter users, Web resources and to browse IMDb movies fingerprints.

9. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *International Conference on User Modeling, Adaptation and Personalization (UMAP), Girona, Spain*. Springer, July 2011.

[2] F. Abel, N. Henze, E. Herder, and D. Krause. Linkage, aggregation, alignment and enrichment of public user profiles with mypes. In A. Paschke, N. Henze, and T. Pellegrini, editors, *Proceedings the 6the International Conference on Semantic Systems, I-SEMANTICS 2010, Graz, Austria, September 1-3, 2010*, ACM International Conference Proceeding Series. ACM, September 2010.

[3] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 971–980, New York, NY, USA, 2007. ACM.

[4] P. Brusilovsky, A. Kobsa, and W. Nejdl, editors. *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*. Springer, 2007.

[5] Z. Chen, J. Cao, Y. Song, Y. Zhang, and J. Li. Web video categorization based on wikipedia categories and content-duplicated open resources. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1107–1110, New York, NY, USA, 2010. ACM.

[21] http://twikime.l3s.uni-hannover.de

Table 6: User study results.

(1^{st} question) - Relevance			(2^{nd} question) - Relatedness			
Agreement	Fingerprint-based(%)	Text-based(%)	Relatedness		Fingerprint-based(%)	Text-based(%)
Strongly Agree	25.74	22.55	Related	5	14.85	15.69
Agree	48.51	32.35		4	29.7	28.43
Undecided	3.96	13.73		3	19.8	17.65
Disagree	12.87	17.65		2	15.84	16.67
Strongly Disagree	8.91	13.73	Unrelated	1	19.8	21.57

[6] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 761–770, New York, NY, USA, 2009. ACM.

[7] M. Grineva, M. Grinev, D. Lizorkin, A. Boldakov, D. Turdakov, A. Sysoev, and A. Kiyko. Blognoon: exploring a topic in the blogosphere. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 213–216, New York, NY, USA, 2011. ACM.

[8] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: search and ranking. In *Proceedings of the 3rd European conference on The Semantic Web: research and applications*, ESWC'06, pages 411–426, Berlin, Heidelberg, 2006. Springer-Verlag.

[9] R. Kawase, P. Siehndel, E. Herder, and W. Nejdl. Hyperlink of men. In *Proceedings of the 2012 Latin American Web Congress (la-web 2012)*, LA-WEB '12, Washington, DC, USA, 2012. IEEE Computer Society.

[10] A. Kittur, E. H. Chi, and B. Suh. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1509–1512, New York, NY, USA, 2009. ACM.

[11] B. Köhncke and W.-T. Balke. Using wikipedia categories for compact representations of chemical documents. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1809–1812, New York, NY, USA, 2010. ACM.

[12] Y. Koren. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 447–456, New York, NY, USA, 2009. ACM.

[13] K. H. Lim and A. Datta. Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media*, MSM '12, pages 25–32, New York, NY, USA, 2012. ACM.

[14] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.

[15] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM.

[16] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[17] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[18] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM Conference on information and knowledge management*, CIKM '07, pages 525–534, New York, NY, USA, 2007. ACM.

[19] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *16th international World Wide Web conference*, New York, NY, USA, 2007. ACM Press.

[20] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 155–162, New York, NY, USA, 2008. ACM.

[21] J. Yu, J. A. Thom, and A. Tam. Ontology evaluation using wikipedia categories for browsing. In *Proceedings of the sixteenth ACM Conference on information and knowledge management*, CIKM '07, pages 223–232, New York, NY, USA, 2007. ACM.