Evidence of Quality of Textual Features on the Web 2.0

Flavio Figueiredo
Dep. of Computer Science
Fed. Univ. of Minas Gerais
flaviov@dcc.ufmg.br

Fabiano Belém
Dep. of Computer Science
Fed. Univ. of Minas Gerais
fmuniz@dcc.ufmg.br

Henrique Pinto
Dep. of Computer Science
Fed. Univ. of Minas Gerais
hpinto@dcc.ufmg.br

Jussara Almeida
Dep. of Computer Science
Fed. Univ. of Minas Gerais
jussara@dcc.ufmg.br

Marcos Gonçalves
Dep. of Computer Science
Fed. Univ. of Minas Gerais
mgoncalv@dcc.ufmg.br

David Fernandes
Dep. of Computer Science
Fed. Univ. of Minas Gerais
davidf@dcc.ufmg.br

Edleno Moura
Dep. of Computer Science
Fed. Univ. of Amazonas
edleno@dcc.ufam.edu.br

Marco Cristo
FUCAPI - Analysis, Research
and Tech. Innovation Center
marco.cristo@gmail.com

ABSTRACT

The growth of popularity of Web 2.0 applications greatly increased the amount of social media content available on the Internet. However, the unsupervised, user-oriented nature of this source of information, and thus, its potential lack of quality, have posed a challenge to information retrieval (IR) services. Previous work focuses mostly only on tags, although a consensus about its effectiveness as supporting information for IR services has not yet been reached. Moreover, other textual features of the Web 2.0 are generally overseen by previous research.

In this context, this work aims at assessing the relative quality of *distinct* textual features available on the Web 2.0. Towards this goal, we analyzed four features (TITLE, TAGS, DESCRIPTION and COMMENTS) in four popular applications (CiteULike, Last.FM, Yahoo! Video, and Youtube). Firstly, we characterized data from these applications in order to extract evidence of quality of each feature with respect to usage, amount of content, descriptive and discriminative power as well as of content diversity across features. Afterwards, a series of classification experiments were conducted as a case study for quality evaluation. Characterization and classification results indicate that: 1) when considered separately, TAGS is the most promising feature, achieving the best classification results, although its absence in a non-negligible fraction of objects may affect its potential use; and 2) each feature may bring different pieces of information, and combining their contents can improve classification.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services

CIKM'09, November 2–6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-512-3/09/11 ...$10.00.

General Terms

Experimentation, Measurement

Keywords

Web 2.0, Textual Features, Social Media

1. INTRODUCTION

Web 2.0 applications have grown significantly in diversity and popularity in recent years. Popular examples include Youtube and Yahoo! Video[1] (or simply YahooVideo), two social video sharing applications, Last.FM[2] (or simply LastFM), an online radio and music community website, and CiteULike[3], a scholarly reference management and discovery service. By distributing mostly *user generated content* and enabling the establishment of online communities and social networks, these applications make use of collaborative knowledge to increase the amount and diversity of content offered. Youtube, for example, is currently the largest video database in the world[4], and the second most searched Website[5]. Although the musical content available in LastFM is not *generated* by its users, it is currently one of the most popular Internet radio stations, due to its community based organization, which gives users the ability to describe musical content and interact over social networks.

Social media is here used to refer to the content, most commonly generated by users, available in Web 2.0 applications. This typically comprises a main object, stored in one of various media types (text, image, audio or video), as well as several other sources of information, commonly in textual form (and thus referred to as *textual features*), associated with it (e.g., tags). Being often unsupervised sources of data, social media offers no guarantee of quality, thus posing a challenge to information retrieval (IR) services such as

[1] `http://youtube.com and http://video.yahoo.com`
[2] `http://last.fm`
[3] `http://www.citeulike.org`
[4] `http://www.comscore.com/press/release.asp?press=1929`
[5] `http://bits.blogs.nytimes.com/2008/10/13/search-ads-come-to-youtube`
URLs were last accessed on June 2009

search, recommendation, and online advertising. In fact, a recent discussion pointed the commonly low quality of multimedia objects, particularly videos, and its impact on the effective use of state-of-the-art multimedia IR techniques as one of the reasons for their limited use on the Web 2.0 [4].

In contrast, tagging has emerged as an effective form of describing and organizing content on the Web 2.0, being a main topic of research in the area. Previous work includes the investigation of its use for supporting search, recommendation, classification and clustering [12, 18, 5, 16] and analyses of tag usage and patterns [7, 19]. However, being also social media content, the effectiveness of tags as supporting information for IR services is questionable. In fact, previous analyses of tags in different applications reach contradictory results [3, 14]. More broadly, other textual features, such as object title and description, do exist in most Web 2.0 applications, but their potential use for IR services have been mostly neglected by previous research.

In this context, this paper aims at providing evidence of the relative quality of *different textual features* commonly found in Web 2.0 applications. The term *quality* is here used to refer to the feature's potential effectiveness as supporting information for IR services. Towards that goal, we crawled object samples from four popular applications, namely, Youtube, YahooVideo, LastFM and CiteULike, collecting the contents of four textual features, namely, TITLE, TAGS, DESCRIPTION and COMMENTS. We then characterized the data in order to extract evidence of quality of each feature with respect to usage, amount of content, descriptive and discriminative power, as well as evidence of content diversity across features. Our main findings are: (1) all four features but TITLE are unexplored in a non-negligible fraction of the collected objects, particularly in applications that allow for the collaborative creation of the feature content; (2) the amount of content tends to be larger in collaborative features; (3) the typically smaller and more often used TITLE and TAGS features exhibit higher descriptive and discriminative power, followed by DESCRIPTION and COMMENTS; (4) there is non-negligible diversity of content across features associated with the same object, indicating that features may bring different pieces of information about it.

We also assessed the relative quality of the features with respect to a specific IR task, that is, object classification, chosen for its common applicability to several services (e.g., automatic construction of web directories) and for enabling automatic assessment of feature quality. To that end, we performed a series of classification experiments, varying the source of information for the classifier and the term weighting scheme. In particular, we tested six classification strategies consisting of taking each feature in isolation as well as two combinations of all four features. Our main results, which are supported by our characterization findings and hold across all weighting schemes analyzed, are: (1) TAGS, if present, are, in fact, the most promising feature in isolation; (2) combining content from all four features can improve the classification results due to the presence of distinct but somewhat complementary content; (3) for classification, feature quality is affected by the amount of content, being TITLE the one with lowest quality, despite its good discriminative/descriptive power and large object coverage.

This paper builds a more solid understanding of issues related to the quality of textual features commonly found in Web 2.0 applications. The knowledge produced here may be of use to Web 2.0 application designers, who can address issues such as the usefulness and attractiveness of each feature to their users. It may also be of relevance to designers of IR services for the Web 2.0 as it uncovers evidence of quality of each feature as a potential source of information. In sum, we here offer a more thorough exploration of the problem than previous work covering different features, in different applications, and thus providing more sound conclusions.

The rest of this paper is organized as follows. Section 2 discusses related work. The textual features analyzed are presented in Section 3, whereas Section 4 summarizes our data collection methodology. Sections 5 and 6 discuss the characterization of the features and the classification experiments, respectively. Finally, conclusions and directions for future work are offered in Section 7.

2. RELATED WORK

Some evidence of quality of Web 2.0 features, mainly tags, has been produced by previous work. Suchanek *et al* [20] look up the amount of known tag terms in Delicious[6] according to lexical databases. They find that approximately half of the used tags have entries in these dictionaries, and that nearly half of these known tags have 10 or more meanings. Bischoff *et al* [3] analyze the quality of tags in the context of search in some Web 2.0 applications. By checking the overlap of tags with content, metadata assigned by experts, and external sources such as search engines, they find that tags tend to be accurate with respect to the object content, and are related to search queries.

In the context classification and clustering, Ramage *et al* [16] contrast the usage of tags assigned by users in Delicious with the usage of the full textual content of the bookmarked webpages. The authors show that combining both webpage content and tags improves the effectiveness of two different clustering algorithms. Another interesting result is that of Chen *et al* [5], which developed a classification model based on both content features and tagging data of musical tracks from LastFM. The authors show that the use of both types of features improves classification results when compared to baseline based only on content features.

In spite these examples, comparing tags with other textual features found on Web 2.0 applications is still an open issue. A very recent publication gives somewhat different insights on the matter, manually comparing three different textual features (TITLE, TAGS and DESCRIPTIONS) when describing versions of images with the same content (i.e., a touristic place), and raising the question whether tags are the most reliable source of information [14]. In comparison with these previous work, we note that our analysis is broader, fully automatic, focusing on far more objects, evaluating many more properties of the features, across different Web 2.0 applications. However, some interesting insights, which can only be captured by human reviews, are presented on that last study. One example are assertions about which features are more semantically related with the context of the image as a whole (i.e., if it is a couple on a holiday or a person sharing the picture with a friend.). The author concludes that tags are not the best textual feature, since title and descriptions better capture the context within the picture and tags are used, mostly, for organizational purposes. This result is contradictory to most previous work on tags.

[6] http://delicious.com

In addition to the analysis of tags, there is also some work on the analysis of commentaries posted by users in blogs [15]. The authors characterize the use, popularity, temporal distribution and amount of information in blog comments. It is shown that the information in comments can be used to increase the amount of documents retrieved by search queries and that the temporal distribution can be use to re-rank query results according to blog entries' post dates.

All of the examples above can be seen as evidence of quality, but they focus either on a single feature or on a small sample of objects. To the best of our knowledge, ours is the first effort to automatically analyze in a large scale the quality of *different textual features* associated with the same object in four different Web 2.0 applications. Moreover, we extract and analyze evidence of quality with regard of usage, quantity, and descriptive and discriminative power.

Recent efforts towards applying multimedia IR techniques on Web 2.0 applications are also worth mentioning. Examples include the use of both content and textual features [17, 5]. These are indications that multimedia IR techniques and social media is converging at some level, a question posed by the previously discussed work of Boll [4]. Also worth mentioning are studies on the matter of social media quality such as the one by Agichtein *et al* [2] focused on the quality of social media objects in QA applications.

3. TEXTUAL FEATURES ON THE WEB 2.0

A Web 2.0 *object* is an instance of a media (text, audio, video, image) in a given Web 2.0 application. There are various sources of information related to an object, here referred to as its features. In particular, *textual features*, the focus of this study, comprise the self-contained textual blocks that are associated with an object, usually with a well defined topic or functionality [6]. We here select four textual features for analysis, namely, TITLE, DESCRIPTION, TAGS, and COMMENTS, which are found in many Web 2.0 applications, although some of them are referred to by different names in some applications. CiteULike, for instance, a social bookmarking website for scientific publications, refers to the DESCRIPTION as ABSTRACT, as it usually contains the publication abstract, and name user COMMENTS as REVIEWS. LastFM refers to COMMENTS as SHOUTS.

Textual features may be categorized according to the level of user collaboration allowed by the application. In particular, the textual features analyzed can be either *collaborative* or *restrictive*. Collaborative features may be altered or appended by any user, whereas restrictive ones may only be altered by one user, typically the one who uploaded the object into the system. This property of a feature is here called its *annotation rights*, a generalization of the previously used *tagging rights* notation [13].

TAGS are a collaborative feature in CiteULike, LastFM and YahooVideo, as any user can add tags to an existing object in these applications. In Youtube, in contrast, only the video uploader can add tags to it. Moreover, while restrictive in both YahooVideo and Youtube, DESCRIPTION has a collaborative nature in both LastFM and CiteULike. In the former, users can edit information on an artist or music in a wiki-like manner. In the latter, users can provide different abstracts to the same publication, although we found that to be very rare in our collected dataset. In all four applications, TITLE is restrictive and COMMENTS is collaborative.

We note that some of the applications may automatically fill some of these features at upload time. Youtube, for instance, adds the name of the object as its TITLE and as TAGS, if these features are not provided. However, it does allow users to remove automatically added TAGS, if a TITLE is provided. In CiteULike, the user may request for the system to extract TITLE and DESCRIPTION from several digital libraries. Nevertheless, in all but one application, users are allowed to change all features, according to their annotation rights. The exception is LastFM, in which TITLE (i.e., artist names) is automatically inserted via media player software based on music files metadata.

4. DATA COLLECTION

In order to perform our study, we built crawlers to sample objects and associated textual features from each application. For Youtube, YahooVideo and LastFM, our crawlers follow a snowball sampling strategy [8]. Each crawler starts with a number of objects as seeds (at least 800). For each object under consideration, it retrieves all related objects (following links provided by these applications), storing them for future consideration. In Youtube and YahooVideo, the seeds were randomly selected from the list of all-time most popular videos provided by the applications. In LastFM, they were selected among the artists associated with the most popular tags[7]. For CiteULike, which does not provide explicit related object links but makes daily snapshots of stored objects available for download, we collected a random sample of one snapshot.

The Youtube, YahooVideo and LastFM crawlers ran for approximately two weeks in July, September, and October 2008, respectively, whereas the sampled CiteULike snapshot is from September 2008. In total, we collected 678,614 articles from CiteULike, 193,457 artist pages from LastFM, and 227,562 and 211,081 videos from YahooVideo and Youtube.

Both Youtube and YahooVideo allow users to assign one of a set of pre-determined categories to their videos. In order to use these categories as object classes in our classification experiments (Section 6), our two crawlers collected the category of each object, along with its associated textual features. For LastFM, even though an official category listing is not provided by the application, external data sources can be used to retrieve the musical genre of artists, which, in turn, can be used as object classes. In particular, as in [5], we sampled musical genres from the AllMusic website[8], which self claims to be the largest musical database available. We were able to find categories for 5,536 artists in our dataset, a larger categorized sample than previous work [5].

Similarly to LastFM, CiteULike does not provide any article category listing. Moreover, we are not aware of any reliable categorization covering multiple scientific areas (as in our sample). Thus, we do not include CiteULike in our current classification experiments, leaving it for future work.

In total, there are 20 and 15 categories in YahooVideo and Youtube, respectively, whereas 9 musical genres were used in the categorization of artists in LastFM. These categories, referred to as object classes, are listed in Table 1.

5. FEATURE CHARACTERIZATION

In this section, we analyze the use of the four selected textual features in the four studied applications. Our character-

[7] Our dataset consists of artist pages including music tracks.

[8] http://www.allmusic.com/

Table 1: Object Classes.

Application	Classes
LastFM	Blues, Classical, Country, R & B, Electronica, Jazz, Pop/Rock, Rap, World
YahooVideo	Action, Animals, Art & Animation, Commercials, Entertainment & TV, Family, Food, Funny Videos, Games, Health & Beauty, How-to, Movies & Shorts, Music, News & Politics, People & Vlogs, Products & Tech., Science & Environment, Sports, Transportation, Travel
Youtube	Autos & Vehicles, Comedy, Education, Entertainment, Film & Animation, Gaming, Howto & Style, Music, News & Politics, Nonprofits & Activism, People & Blogs, Pets & Animals, Science & Technology, Sports, Travel & Events

ization, performed over our collected datasets[9], covers four aspects, aiming at providing evidence of the relative quality of the features as supporting information for IR tasks.

First, we characterize feature *usage*, investigating which features are more explored by users (Section 5.1). Feature usage is of key importance, as under-explored features, that is, features that are absent in a non-negligible fraction of the objects, may not be a reliable source of information. Second, we characterize the *amount of content* in each feature (Section 5.2). If present, a feature should also provide enough content to be effective for IR. Third, we use heuristics, adapted from previous work [6], to assess the *descriptive and discriminative power* of each feature (Section 5.3). Effective sources of information for IR tasks should offer a reasonably accurate description of the object content and/or discriminate objects into different pre-defined categories (for services such as browsing, recommendation and advertising) and into levels of relevance (for searching). Lastly, we characterize the *diversity of content* across features (Section 5.4), motivating the exploration of different feature combination strategies in the next section. The main findings from our characterization are summarized in Section 5.5. Throughout this section, we refer to the set of terms in a feature F associated with one given object as an *instance* of feature F.

5.1 Feature Usage

Table 2 shows, for each feature and application, the percentage of empty feature instances, i.e., objects with no terms in the given feature. The annotation rights of each feature is shown in parenthesis, C for collaborative and R for restrictive. The fraction of empty instances is much larger for collaborative features. In fact, some of these features, such as COMMENTS in all applications but Youtube, and DESCRIPTION in LastFM and CiteULike, are vastly under-explored. Even Youtube, with millions of users, has a significant fraction of empty COMMENTS, similarly to previous findings on blog and photo sharing systems [15, 14]. The TAGS feature, focus of most previous efforts to enhance Web 2.0 IR services, is also absent in non-negligible fractions of 16% and of almost 19% of the objects collected from YahooVideo and LastFM, respectively. Only TITLE, restrictive in all applications, is present in practically all objects.

Table 2: Percentage of Empty Feature Instances (C = collaborative feature, R = restrictive feature).

	TITLE	TAGS	DESC.	COMM.
CiteUL	0.53% (R)	8.26% (C)	51.08% (C)	99.96% (C)
LastFM	0.00% (R)	18.88% (C)	53.52% (C)	54.38% (C)
Yahoo	0.15% (R)	16.00% (C)	1.17% (R)	96.88% (C)
Youtube	0.00% (R)	0.06% (R)	0.00% (R)	23.36% (C)

These results may be partially explained by restrictions imposed by the application, such as the case of TITLE and TAGS in Youtube, which may be automatically filled by the system. However, none of the applications enforces the usage of DESCRIPTION, for instance. Nevertheless, it has a much larger presence in YahooVideo and Youtube, where it is restrictive, than in the other two applications. In fact, in YahooVideo, it has a much larger presence than TAGS, a collaborative feature. This is surprising, as we expected that users would be more willing to provide a few words as TAGS than a few sentences as DESCRIPTION. An interesting comparison is that of TAGS in CiteULike and LastFM. Even though both applications motivate the organization of personal libraries, CiteULike does explicitly ask users to associate TAGS[10] to their articles, whereas no such incentive exists in LastFM. Comparing the usage of TAGS in both applications, it seems like users may need extra incentives to use collaborative features, an issue previously raised for tagging systems [13].

Thus, considering solely the object coverage in terms of feature presence, the restrictive TITLE offers the best quality of all features, in all applications. This is consistent with a previous analysis of Flickr[11], which also reported a much larger presence of TITLE than of TAGS and DESCRIPTION [14]. More broadly, the feature annotation rights seems to play an important role on its usage. In particular, using only collaborative TAGS, DESCRIPTION or COMMENTS as single source of data for IR tasks may not be effective due to the lack of information for a significant fraction of objects.

5.2 Amount of Content

In this section, we analyze the amount of content available in textual features. Since this analysis is language-dependent, we focus on the English language, applying a simple filter that disregards objects with less than three English stop-words in their textual features[12]. We also focus our analysis on *non-empty feature instances*.

After removing non-english objects and empty feature instances, we were left with a varying number of objects for the analysis of each feature in each application. This number is larger than 150,000 for TITLE, DESCRIPTION and TAGS in all applications but LastFM, for which they exceeded 86,500 objects. Our cleaned datasets also include 152,717, 76,627, 6,037 and 76 objects with COMMENTS in Youtube, LastFM, YahooVideo and CiteULike, respectively. We removed terms containing only non-alphanumeric characters, and applied the Porter stemming algorithm[13] to remove affixes, remov-

[9] The complete LastFM dataset, with 193,457 objects, was used.

[10] We note that "no-tag" entries, automatically added by the system, are considered as empty instances in our analysis.

[11] `http://www.flickr.com.`

[12] `http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words`

[13] `http://tartarus.org/~martin/PorterStemmer/`

Table 3: Vocabulary Size of Non-Empty Feature Instances.

	TITLE				TAGS				DESCRIPTION				COMMENTS			
	AR	μ	CV	max	AR	μ	CV	max	AR	μ	CV	max	AR	μ	CV	max
CiteULike	R	7.5	0.40	73	C	4.0	1.33	194	C	65.2	0.5	3890	C	51.9	1.42	449
LastFM	R	1.8	0.47	23	C	27.4	1.49	269	C	90.1	1.06	3390	C	110.2	3.55	22634
YahooVideo	R	6.3	0.39	16	C	12.8	0.52	52	R	21.6	0.71	141	C	52.2	2.51	4189
Youtube	R	4.6	0.43	36	R	10.0	0.60	101	R	40.4	1.75	2071	C	322.3	1.94	16965

ing stop-words at the end. This was done to remove meaningless terms and normalize semantically equivalent ones.

We characterize the *number of distinct terms* in each feature instance, here referred to as its *vocabulary size*, which represents the amount of content available for use by IR services. These results are summarized in Table 3, which presents the mean μ, coefficient of variation CV (ratio of standard deviation to the mean), and maximum values[14]. The AR columns show the feature annotation rights.

In general, TITLE has instances with the smallest vocabulary, followed by TAGS, DESCRIPTION and COMMENTS. Moreover, with few exceptions, instances of collaborative features tend to have vocabularies with larger average sizes and high variability (CV). For instance, the table shows that COMMENTS instances (if present) have the largest vocabulary, on average (except in CiteULike), also exhibiting very large coefficient of variation. Similarly, instances of DESCRIPTION in CiteULike and LastFM, and of TAGS in LastFM, also collaborative, have larger vocabularies than instances of the same features in Youtube, where they are restrictive. Nevertheless, COMMENTS carry much less content in YahooVideo and CiteULike than in the other two applications, as it is not often explored by their users (Section 5.1). These results are consistent with the ones previously observed for TITLE, TAGS and DESCRIPTION in Flickr [14].

In CiteULike, the TAGS feature, in spite of being collaborative, tends to have instances with smaller vocabularies, providing less content than TITLE, unlike in the other applications. This is possibly due to long article titles and the stabilization of tagging vocabulary with time [7, 12]. Nevertheless, a small fraction of objects have TAGS with very large vocabularies (up to 194 terms). Unlike in all other applications, DESCRIPTION has larger instances on average, possibly because they typically carry article abstracts, with more information than the poorly used COMMENTS.

In LastFM, TITLE has a much stronger bias towards small vocabularies than in the other applications. This is expected since the feature is usually referring to artist names. In contrast, the TAGS feature tends to carry much more content, with an average of 27 terms and reaching up to a maximum of 269. COMMENTS is the feature with the largest amount of content, but this average is biased towards very popular artists, for which COMMENTS can have more than 22,000 terms. DESCRIPTION instances also have large vocabularies, larger than in the other applications, likely due to its collaborative nature, as users collectively write a text, in a wiki-like manner, describing the artists.

YahooVideo and Youtube show similar trends, although the vocabularies of instances of both DESCRIPTION and COMMENTS tend to be much larger in Youtube, possibly due to its larger audience. In contrast, the instances of TITLE and TAGS have larger vocabularies in YahooVideo. Whereas the differences in TAGS may be due to the higher degree of collaboration in YahooVideo, the differences in TITLE may reflect differences in usage patterns.

5.3 Descriptive and Discriminative Power

Web 2.0 features normally define different content blocks in Web pages. For instance, in Youtube, the comments about a video are placed near to each other into a common (comments) block. Thus, in this section, we use metrics previously proposed to assert the importance of traditional Web page blocks to assess the importance of the four textual features analyzed. In particular, we adapt metrics based on heuristics, proposed as part of an Information Retrieval model [6], to the context of textual features of Web 2.0 objects in order to assess their descriptive and discriminative power. In the following, we first formally define each metric, and then present the characterization results.

5.3.1 Descriptive Power

Our assessment of the descriptive power of each feature is based on a *heuristic* metric called *Average Feature Spread*, which can be computed as follows. We start by defining the spread of a term in a given object. Let o be an object in the collection O, and t a term which appears in at least one feature instance f associated with o. The *term spread*, $TS(t, o)$, measures the number of feature instances associated with o which contain t, that is:

$$TS(t,o) = \sum_{f \in o} i, \text{ where } i = \begin{cases} 1 & \text{if } t \in f \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

The assumption behind $TS(t, o)$ is that the larger the number of features of o that contain t, the more related t is with o's content. For instance, if the term "Sting" appears in all features of a given video, there is a high chance that the video is related to the famous singer.

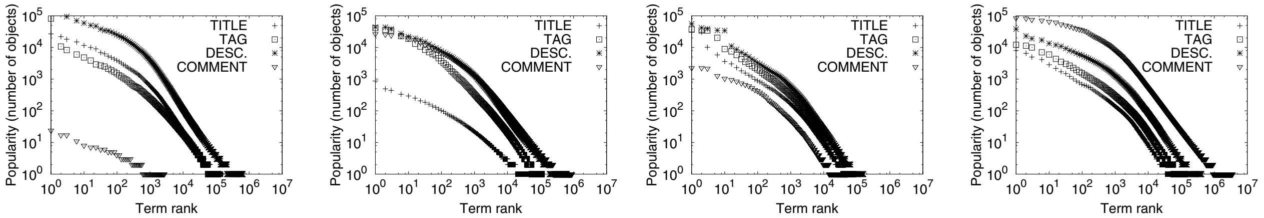
Next, we define the *Feature Instance Spread* of a feature instance f associated with an object o, $FIS(f, o)$, as the average term spread across all terms in f. That is, given $|f|$ the number of terms in instance f, $FIS(f, o)$ is defined as:

$$FIS(f,o) = \sum_{t \in f} \frac{TS(t,o)}{|f|}. \qquad (2)$$

The *Feature Instance Spread* is a heuristic to assess how the terms of a given feature instance f are related to the content of instances of other features associated with the same object o. It is thus a heuristic to estimate how the feature instance f is related to o's content, being here used as an estimate of the average descriptive power of feature instance f for the given object o.

Following this reasoning, the descriptive power of a feature F in the object collection O can be captured by averaging the values of FIS across all objects. We refer to this metric

[14] The probability distributions lead to similar conclusions as the aggregated results in Table 3 and were omitted for space reasons.



(a) CiteULike (b) LastFM (c) YahooVideo (d) Youtube

Figure 1: Term Popularity Distributions.

Table 4: *AFS*, *AIFF* and *FI* Values.

	CiteULike			LastFM			YahooVideo			Youtube		
	AFS	*AIFF*	*FI*	*AFS*	*AIFF*	*FI*	*AFS*	*AIFF*	*FI*	*AFS*	*AIFF*	*FI*
TITLE	1.916	11.773	22.557	2.653	10.723	28.448	2.262	10.724	24.258	2.536	11.227	28.472
TAGS	1.629	11.635	18.953	1.328	10.112	13.429	1.868	10.495	19.605	2.073	11.135	23.083
DESCRIPTION	1.126	11.953	13.459	1.211	10.583	12.816	1.511	10.777	16.284	1.721	11.259	19.377
COMMENTS	-	-	-	1.207	10.623	12.822	-	-	-	1.124	11.517	12.945

as the *Average Feature Spread* of F, $AFS(F)$. Given each object o in the collection and the instance of F associated with it, f, the $AFS(F)$ is computed as:

$$AFS(F) = \frac{\sum_{o \in O} FIS(f, o)}{|O|} \tag{3}$$

5.3.2 Discriminative Power

In order to assess the *discriminative power* of a feature, we use a heuristic called *Average Inverse Feature Frequency (AIFF)*. The $AIFF$ metric builds on a small variation of the IDF metric, called *Inverse Feature Frequency (IFF)*, which considers instances of a feature F as a separate "document collection". Given a feature F with $|F|$ instances[15], and a term t that occurs in at least one of F's instances, the $IFF(t, F)$ of term t in F is defined as:

$$IFF(t, F) = \log(\frac{|F|}{Frequency(t, F)}), \tag{4}$$

where $Frequency(t, F)$ is the number of instances of F in which the term t appears.

The IFF metric assesses how much information carries *the occurrence of a given term in a given feature.* The assumption is that terms occurring in many instances of the feature are bad discriminators of content. For example, whereas the occurrence of "music" in a TITLE of a Youtube object brings little information about its content in relation to other music videos, the occurrence of "Sting" may be more useful to discriminate it from the other objects.

We then define the *Average Inverse Feature Frequency* of feature F, $AIFF(F)$ as the average IFF over all terms that occur in all instances of F. It is thus a heuristic to estimate the discriminative power of the feature F in the object collection. Given $|V_F|$ the size of the complete vocabulary of feature F (i.e., considering all instances of F in the object collection), the $AIFF(F)$ is computed as:

$$AIFF(F) = \frac{\sum_{t \in F} IFF(t, F)}{|V_F|} \tag{5}$$

Finally, the values of FS and $AIFF$ are used to estimate the *Feature Importance* (FI), which measures how good a feature is considering both its discriminative and descriptive power. The FI of a feature F is given by their product, that is $FI(F) = AFS(F) \times AIFF(F)$

5.3.3 Results

We computed the $AIFF$, AFS and FI values of each feature in all four applications, using our stemmed datasets and considering only objects with non-empty instances of *all* features. Given the negligible fractions of non-empty COMMENTS in YahooVideo and CiteULike, we disregarded this feature in both applications. As Table 4 shows, the AFS values provide a consistent ranking of features in all four applications. TITLE is the most descriptive feature, followed by TAGS, DESCRIPTION and, if considered, COMMENTS. In contrast, the $AIFF$ values show little distinction across features in any application. The FI metric, dominated by its AFS component, produces a ranking in agreement with it.

In order to understand why the $AIFF$ metric is not able to clearly distinguish one feature from the other, we plotted the term popularity distribution of each feature, considering all instances of the feature. Figure 1 shows that the distributions are heavy-tailed in all applications, thus containing a large number of terms with very low popularity. These terms have very large IFF values, and end up boosting the $AIFF$ of all features somewhat similarly. In fact, previous studies [16, 9] showed that IDF, on which $AIFF$ is based, over-emphasizes unique and very unpopular terms, and that these terms are not helpful for tasks such as clustering and classification, since, being rarely shared, they are unable to help grouping objects into semantically meaningful groups.

Thus, we recomputed the $AIFF$ values considering only terms that appeared in more than 50 instances[16]. The re-

[15] We note that $|F|$ is not necessarily equal do $|O|$, since some objects may have empty instances of F.

[16] Other thresholds achieve similar results (10, 100 and 1000).

Table 5: $AIFF$ Values (ignoring unpopular terms).

	CiteULike	LastFM	YahooVideo	Youtube
TITLE	7.311	6.644	6.674	7.126
TAGS	7.595	5.997	6.548	6.997
DESC.	7.028	5.833	6.371	6.739
COMM.	-	5.908	-	6.649

sults, presented in Table 5, show a more clear distinction between the features. Overall, according to the $AIFF$ heuristic, TITLE is the most discriminative feature, followed by TAGS, DESCRIPTION and, if considered, COMMENTS, a ranking that is consistent with the AFS results. The exception is CiteULike in which TAGS has a larger $AIFF$ value than TITLE. This may be due to the use of terms in article titles that are less specific than those in TAGS, possibly to facilitate discovery and understanding of the article subject by potential readers. As consequence, the same terms may co-occur in many TITLE instances, thus reducing its IFF.

5.4 Content Diversity

So far we have characterized aspects of each feature separately. We now investigate whether different features contribute with different content (i.e., terms) about the associated object. As in Section 5.2, we focus on the English language, using the stemmed datasets. We quantify the *similarity* between two feature instances in terms of term co-occurrence using the Jaccard Coefficient. Given two sets of items T_1 and T_2, the Jaccard Coefficient is computed as $J(T_1, T_2) = \frac{|T_1 \bigcap T_2|}{|T_1 \bigcup T_2|}$.

We compute the similarity between two features associated with the same object using as input sets the N most highly ranked terms from each feature instance based on the product of the TS and IFF metrics[17]. We then compute the average similarity between two features across all objects containing non-empty instances of them.

Table 6 shows the average similarity between all pairs of features in all four applications for N=5 and, in parenthesis, when all terms of each feature instance are considered. Results for N=15, 30 and 50 are in the same range. As in the previous section, we disregard COMMENTS in CiteULike and YahooVideo. There seems to be more similarity between restrictive features (e.g., TITLE and DESCRIPTION in YahooVideo and Youtube), as the same user tends to use common words in them. The exception is TAGS in YahooVideo, which, despite collaborative, shares some similarity with the restrictive TITLE and DESCRIPTION features, perhaps an indication that TAGS are often used only by the video owner herself. Nevertheless, the Jaccard Coefficients are all under 0.52. Thus, although there is some co-occurrence of highly ranked terms across features, each feature may still bring new (possibly relevant) information about the object.

Table 6: Average Similarity (Jaccard Coefficient) Between Non-Empty Feature Instances.

	CiteULike	LastFM	Yahoo	Youtube
TITLE×TAGS	0.13 (0.09)	0.07 (0.01)	0.52 (0.33)	0.36 (0.25)
TITLE×DESC.	0.31 (0.09)	0.22 (0.03)	0.40 (0.20)	0.28 (0.015)
TAGS×DESC.	0.13 (0.02)	0.13 (0.04)	0.43 (0.20)	0.32 (0.14)
TITLE×COMM.		0.12 (0.02)		0.14 (0.02)
TAGS×COMM.		0.10 (0.03)		0.17 (0.02)
DESC.×COMM.		0.18 (0.03)		0.16 (0.03)

5.5 Summary

Our characterization results may be summarized into four main findings. First, all four features but TITLE, have a non-negligible fraction of empty instances in at least two of the analyzed applications, and thus may not be effective as single source of information for IR services. More broadly, restrictive features seem to be more often explored by users than collaborative ones, even within the same feature category. Second, the amount of content tends to be larger in collaborative features. Third, the typically smaller and more often used TITLE and TAGS exhibit higher descriptive and discriminative power, according to our AFS and $AIFF$ heuristics, followed by DESCRIPTION and COMMENTS. Finally, there is significant content diversity across features associated with the same object, indicating that each feature may bring different pieces of information about it.

In the next section, we assess the relative quality of the textual features when applied to a specific IR task, namely, object classification. We choose to focus on classification, leaving the analysis of other IR tasks for future work, for its common applicability to several services such as automatic construction of web directories and recommendation, as well as for allowing the automatic assessment of feature quality. The classification results are discussed in light of our characterization findings, as amount of content, descriptive and discriminative power have impact on the effectiveness of the feature as source of information for object classification. Moreover, the content diversity observed motivates the experimentation with feature combination strategies.

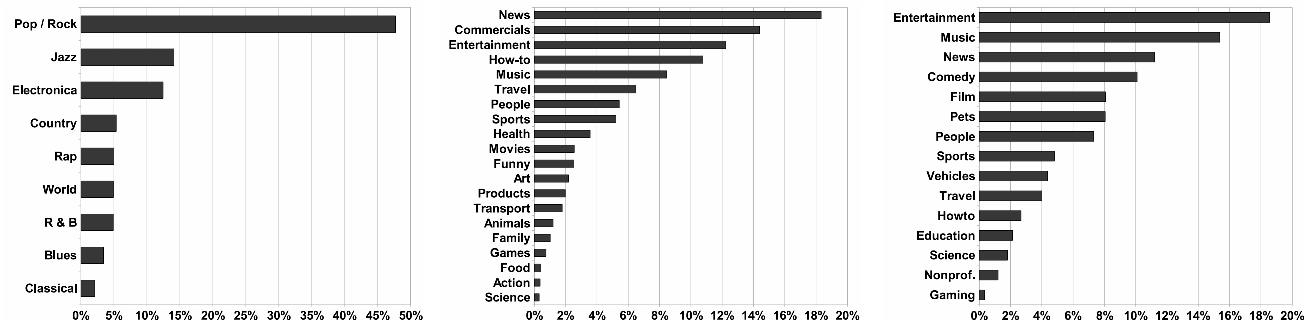
6. OBJECT CLASSIFICATION

This section presents our object classification experiments, which use the collected categories, introduced in Section 4, as object labels. As discussed in that section, we focus our experiments on LastFM, Youtube and YahooVideo. In Section 6.1, we present the model adopted for object representation, and define the various classification strategies considered. Our experimental setup is described in Section 6.2, whereas the main results are discussed in Section 6.3.

6.1 Object Representation Model

In order to perform object classification, we first need to define a model to represent the objects. We adopt the vector space model (VSM), representing each object as a vector in a real-valued space whose dimensionality |V| is the size of the vocabulary of the object feature(s) being considered as information source. Each term in the feature(s) is represented by a real value in the vector, which must, ideally, represent the degree of relationship between the term and the object it was assigned to. Once defined the object representation model, two important issues are: (1) how to determine the weight of each term in order to capture the semantics of the object, and (2) how to model the objects in the VSM using their distinct textual features. We considered the following term weighting schemes:

[17] Qualitatively similar results are obtained using $TF \times IFF$.



(a) LastFM

(b) YahooVideo

(c) Youtube

Figure 2: Distribution of Objects Across Classes.

TS : the weight of a term t in an object o is equal to the spread of t in o, $TS(t, o)$, which heuristically captures the descriptive power of the term (see Section 5.3.1).

TS × IFF : the weight of t is equal to the product $TS(t, o) \times IFF(t, F)$, capturing, thus, both descriptive and discriminative powers of t. F represents the feature (or feature combination) used to represent the object (see below).

TF : the weight of t is given by the frequency of t in the feature (combination) used to represent the object.

TF × IFF : the weight is given by the product $TF \times IFF$.

The last two schemes allow us to compare the TS and the more commonly used Term Frequency (TF) metrics.

We also examine six strategies to model an object o as a vector V, which is normalized so that $||V|| = 1$:

Title : only terms present in the TITLE of o constitute the vector V, that is, $V = V_{title.}$.

Tags : $V = V_{tags}$, where V_{tags} is the vector that represents only terms in the TAGS of o.

Description : $V = V_{desc.}$, where $V_{desc.}$ is the vector that represents only terms in the DESCRIPTION of o.

Comments : $V = V_{comm}$, where V_{comm} is the vector that represents only terms in the COMMENTS of o.

Bag of Words : all four features are taken as a single document, as done by traditional IR algorithms, which do not consider the block structure in a Web page. Therefore, the same terms in different features are represented by the same element in the vector.

Concatenation All textual features are concatenated in a single vector, so that the same terms in different features are considered different elements in the vector, i.e., vector $V_{conc.}$ is given by $\langle V_{comm.}, V_{desc.}, V_{tag}, V_{title} \rangle$.

In each strategy, vector V is defined as $\langle w_{f1}, w_{f2}, ..., w_{fn} \rangle$ where w_{fi} is the weight of term t in the instance of the considered feature f. In particular, vector V_{bagw} is defined as $\langle w_{o1}, w_{o2}, ..., w_{on} \rangle$ where w_{oi} is the weight of term t within the entire object. Note that, in this case, the IFF metric is equal to the more traditional IDF metric.

These strategies do not cover all possible combinations of features, but are useful to compare their quality for object classification. In particular, the last two strategies are motivated by the results in Section 5.4, and allow us to investigate how textual features that are effective, when used in isolate, compare with the combination of multiple features.

6.2 Experimental Setup

Our classification experiments were performed using a Support Vector Machine (SVM) algorithm with linear kernel implemented in the Liblinear tool [1]. This algorithm was selected because it is an efficient state-of-art classification algorithm for large samples of textual data, and because linear kernels work particularly well for text classification tasks [11]. As before, we used our stemmed datasets, considering only objects with non-empty instances of all features and classes associated, except in YahooVideo, for which the COMMENTS feature was disregarded. Moreover, as Figure 2 shows, some object classes are highly underpopulated in our datasets[18]. Thus, we choose to filter out all classes with a fraction of objects under than 2.2%, removing 8 and 4 classes from YahooVideo and Youtube, respectively.

Our experiments consisted of 10 runs, each with a distinct sample of 5000 objects from each application, using 10-fold cross-validation within each sample. Best SVM parameters (i.e., cost C) were searched for within each training sample, being the default (C=1) the best one in most cases.

We assess the quality of the features using two commonly used classification evaluation metrics, namely *Macro-F1* and *Micro-F1*. These metrics capture both precision and recall of the classification. Let TP, FP and FN be, respectively, the numbers of true positives, false positives and false negatives of the classification output for a class c. *Precision* is defined as the fraction of correctly classified objects of a class in relation to all objects that were assigned to that class, i.e., $Precision = \frac{TP}{TP+FP}$. On the other hand, *Recall* is the fraction of objects of a class that were assigned to that class by the classifier, that is, $Recall = \frac{TP}{TP+FN}$.

Precision and *Recall* are combined into the $F1$ metric such that $F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$. Overall results for all classes are provided by *Macro-F1* and *Micro-F1*. *Macro-F1* is defined as the average of the $F1$ values over all classes,

[18] The class names in Figure 2 are reduced due to space constraints. We refer to Table 1 for their complete names.

Table 7: Classification Results: Macro-F1 (Average and 90% Confidence Intervals).

Object Model	LastFM	YahooVideo	Youtube
Title	0.204±0.016	0.522±0.005	0.404±0.004
Tag	0.809±0.015	0.635±0.004	0.540±0.004
Description	0.752±0.016	0.574±0.005	0.437±0.004
Comment	0.528±0.020		0.464±0.004
Bag of Words	0.807±0.013	0.663±0.004	0.594±0.004
Concatenated	0.805±0.015	0.665±0.004	0.567±0.004

Table 8: Classification Results: Micro-F1 (Average and 90% Confidence Intervals).

Object Model	LastFM	YahooVideo	Youtube
Title	0.520±0.013	0.569±0.003	0.434±0.003
Tag	0.866±0.007	0.671±0.003	0.562±0.003
Description	0.831±0.009	0.636±0.003	0.470±0.003
Comment	0.721±0.014		0.516±0.004
Bag of Words	0.870±0.008	0.705±0.003	0.628±0.004
Concatenated	0.868±0.009	0.709±0.003	0.600±0.003

whereas *Micro-F1* is computed using values of TP, FP and FN calculated as the sum of respective values for all classes.

6.3 Results

We performed classification experiments combining each term weighting scheme with each feature or feature combination object model. Initially we present the results obtained using the $TS \times IFF$ weighting scheme, as these are the heuristics proposed for assessing the descriptive and discriminative power of featureS. Tables 7 and 8 show the Macro-F1 and Micro-F1 values for each feature combination strategy.

Considering the results obtained when each feature is used in isolation as object representation, TAGS are, undoubtedly, the best single feature in all applications, considering both Micro and Macro-F1 values. This is consistent with our characterization results which show that TAGS have: (1) good descriptive and discriminative power, according to our heuristics, with AFS and $AIFF$ values close to those of TITLE, and (2) at least twice more terms than TITLE, on average, in the three applications. The larger amount of content favors TAGS as source of information for the classifier, as SVM is known to work better in the presence of larger information (term) spaces [11]. This issue of amount of content may also explain the poor performance of TITLE, the worst feature in all applications, despite its AFS and $AIFF$ values being the largest ones. For example, instances of TITLE in LastFM typically contain the name of the artist, which is usually very short. Regarding the other two features, COMMENTS is better than DESCRIPTION in Youtube, considering both metrics. In spite of the smaller AFS (descriptive power) and a somewhat similar $AIFF$ (discriminative power), COMMENTS instances have, on average, more than 8 times more terms than DESCRIPTION instances, which turns to be a dominant factor for the classification effectiveness. In contrast, we find that DESCRIPTION outperforms COMMENTS in LastFM, in spite of the somewhat larger amount of content, AFS and $AIFF$ of the latter. Despite not completely supported by our characterization, this result may reflect the wiki-like collaborative nature of the DESCRIPTION feature in LastFM, which brings a tendency for a high quality semantic content, a phenomenon also observed in Wikipedia [10]. This is an aspect involving social behavior not captured by our current metrics, and is subject of future work.

The Jaccard Coefficients presented in Section 5.4 suggests that there may exist distinct content in each feature which can be leveraged for classification purposes. This is confirmed by the results of both feature combination strategies.

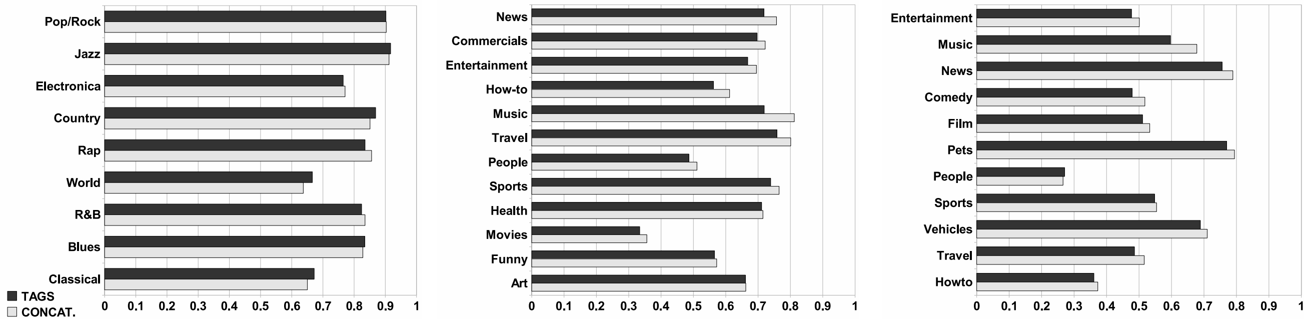
The tables also show that, in most cases, for a given feature (combination) strategy, the classification results in LastFM, for both Micro and Macro-F1, are much higher than in the other applications. This is possibly because LastFM object classes are defined by experts of the music industry, guaranteeing a higher level of consistency in the class assignments when compared to the other applications, in which individual users are responsible for the manual classification. This problem is exacerbated if we consider that Youtube and YahooVideo have a larger number of classes, and some of them may have semantic overlap, making it even harder for users to determine to which class an object belongs. For example, a comedy video can be rightfully inserted into either the "Entertainment" or the "Funny Videos" category. Figure 3 shows the results of F1 values per class for both Tags and Concatenation strategies. Clearly, the classifier achieves higher F1 values in LastFM, being greater than 0.65 for all classes. Moreover, the results are more evenly distributed across classes. In contrast, in Youtube and YahooVideo, the F1 values can be as low as 0.27 and 0.33, being also more unevenly distributed.

Finally, we analyzed the impact of the different term weighting schemes, and found that, surprisingly, results do not differ significantly. This finding is, at first sight, not in consonance with previous work [16, 9] which showed that weighting schemes based on IDF-like metrics may be detrimental to the effectiveness of some IR services because noisy and non-informative terms. Nevertheless, we should consider that SVM is robust to such terms, being capable of filtering some of them out in a feature selection manner [11].

In sum, we can conclude that: 1) TAGS, when present, is the best feature in isolation for classification purposes due to a combination of good descriptive and discriminative power and large amount of content; 2) combination of features may bring some benefits due to the presence of distinct but somewhat complementary content; 3) terms in different feature instances should be considered in their original context in order to preserve their descriptive/discriminative power; 4) a combination of fewer classes, with less ambiguity, and more qualified human labelers, made classification more effective in LastFM than in the video applications; and 5) the use of different weighting schemes does not produce much difference due to inherent properties of the SVM classifier.

7. CONCLUSIONS AND FUTURE WORK

We investigated issues related to the quality of textual features in Web 2.0 applications. To provide insights on the matter, we sampled data from TITLE, DESCRIPTION, TAGS and COMMENTS associated with objects in CiteULike, LastFM, YahooVideo and Youtube. Our characterization of the features revealed that collaborative features, including TAGS, are absent in a non-negligible fraction of objects. However, if present, they tend to provide more content than restrictive features, such as TITLE. We also found that the smaller TITLE and TAGS have higher descriptive and discriminative power, according to our adapted heuristics, and that there exists significant content diversity across features. As



(a) LastFM

(b) YahooVideo

(c) Youtube

Figure 3: F1 measures per class for Tags and Concatenation

a case study, we also assessed the quality of features when applied to object classification. Our results show that TAGS, if present, is the most promising feature. However, its absence in a non-negligible fraction of objects in some applications should be of concern. Moreover, we found that combining content from all features can lead to classification improvements. Though widely present and highly ranked according to both descriptive and discriminative heuristics, the TITLE feature achieved the worst classification results, being severely impacted by the small amount of content.

Future work includes further investigation of feature quality issues in classification and other IR services, correlating our results with social network aspects, and exploring feature quality enhancement strategies.

8. ACKNOWLEDGEMENTS

Flavio Figueiredo and Jussara Almeida are sponsored by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process number 20090215103600. This research was also partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT / CNPq / INCT Web Grant Number 573871/2008-6), by project REBU (MCT / CNPq / CTInfo Grant Number 550995 / 2007-2) and by project Infoweb (MCT / CNPq / CTInfo Grant Number 55.0874 / 2007-0).

9. REFERENCES

[1] Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding High-Quality Content in Social Media. In *Proc. WSDM*, 2008.

[3] K. Bischoff, F. Claudiu-S, N. Wolfgang, and P. Raluca. Can All Tags Be Used for Search? In *Proc. CIKM*, 2008.

[4] S. Boll. MultiTube–Where Web 2.0 and Multimedia Could Meet. *IEEE Multimedia*, 14(1), 2007.

[5] L. Chen, P. Wright, and W. Nejdl. Improving music genre classification using collaborative tagging data. In *Proc. WSDM*, 2009.

[6] D. Fernandes, E. de Moura, B. Ribeiro-Neto, A. da Silva, and M. Gonçalves. Computing Block Importance for Searching on Web Sites. In *Proc. CIKM*, 2007.

[7] S. Golder and B. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 2006.

[8] L. A. Goodman. Snowball Sampling. *Annals of Math. Statistics*, 32(1), 1961.

[9] T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *Proc. WWW*, 2002.

[10] M. L. E. Hu, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proc. CIKM*, 2007.

[11] T. Joachims, C. Nedellec, and C. Rouveirol. Text categorization with support vector machines: learning with many relevant. In *Europ. Conf. on Machine Learning*. Springer, 1998.

[12] X. Li, L. Guo, and Y. Zhao. Tag-based Social Interest Discovery. In *Proc. WWW*, 2008.

[13] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, Toread. In *Collaborative Web Tagging Workshop (WWW'06)*, 2006.

[14] C. Marshall. No Bull, No Spin: A comparison of tags with other forms of user metadata. In *Proc. JCDL*, 2009.

[15] G. Mishne. Using blog properties to improve retrieval. *Proc. of ICWSM*, 2007.

[16] D. Ramage, P. Heymann, C. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proc. WSDM*, 2009.

[17] M. Rege, M. Dong, and J. Hua. Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering. In *Proc. WWW*, 2008.

[18] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. Parreira, and G. Weikum. Efficient Top-k Querying Over Social-Tagging Networks. In *Proc. SIGIR*, 2008.

[19] B. Sigurbjornsson and R. van Zwol. Flickr Tag Recommendation Based on Collective Knowledge. In *Proc. WWW*, 2008.

[20] F. Suchanek, M. Vojnovic, and D. Gunawardena. Social Tags: Meanings and Suggestions. In *Proc. CIKM*, 2008.