JIS

Article

Topical tags vs. non-topical tags: towards a bipartite classification?

Journal of Information Science
1–24

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Valerio Basile
Center for Language and Cognition, University of Groningen, Groningen, The Netherlands

Silvio Peroni
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy

Fabio Tamburini
Department of Classic Philology and Italian Studies, University of Bologna, Bologna, Italy

Fabio Vitali
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Abstract
In this paper we investigate whether it is possible to create a computational approach that allows us to distinguish topical tags (i.e., talking about the topic of a resource) and non-topical tags (i.e., describing aspects of a resource that are not related to its topic) in folksonomies, in a way that correlates with humans. Towards this goal, we collected 21M tags (1.2M unique terms) from Delicious and we developed an unsupervised statistical algorithm that classifies such tags by applying a word space model adapted to the folksonomy space. Our algorithm analyses the co-occurrence network of tags to a target tag and exploits graph-based metrics for their classification. We validated its outcomes against a reference classification made by humans on a limited number of terms in three separate tests. The analysis of the outcomes of our algorithm shows, in some cases, a consistent disagreement among humans and between humans and our algorithm about what constitutes a topical tag, and suggests the rise of a new category of overly generic tags (i.e., umbrella tags).

Keywords
Delicious; folksonomy; latent semantic analysis; topicality and non-topicality of tags; umbrella tags; user testing session

1. Introduction

Folksonomies (a portmanteau of *folk* and *taxonomy*) [1] are light-weight semantic artefacts built by end users of the Social Web through the simple mechanism of associating tags – i.e., arbitrary strings not restricted to a given vocabulary – to resources such as text documents (e.g. Mendeley), pictures (e.g., Flickr), songs (e.g., Last.fm), Web bookmarks (e.g. Delicious), library items [2], and really any old *thing* [3]. The actual meaning of a particular tag depends on multiple factors: the user who tagged the entity, his/her social context, the temporal context of the tagging, the domain covered by the folksonomy in consideration, etc.

Along the line of the Web Science vision [4], which considers the Web a social artefact to be studied through cognitive and scientific tools, several works have appeared studying the use of tags in folksonomic collections. In addition, folksonomies and the implicit relations that can be derived by their analysis, e.g., those connecting users on the base of similar behaviours and interests, can have consequences on other technologies, for instance recommendation systems [5]. In this domain, folksonomic tags – and, in particular, the classification of tags according to different

Corresponding author:
Silvio Peroni, Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni 7, 40126 Bologna (BO), Italy
silvio.peroni@unibo.it

taxonomies, e.g., content-related vs. audience-related tags – were and are actively used for increasing the effectiveness of content-based recommender systems, collaborative filtering systems, and social recommender systems [6].

One of the most common requirements, thus, is to find out the reason for which folksonomic tags were created. Most tags are meant to describe the content of the associated resource, but, as Kipp points out, "Previous studies of social tagging systems [...] all report that while most tags are subject related, there is often a small but significant core of tags which are not subject related at all" [7]. Yet, while faceted or ontological approaches to classification would insist on providing the rationale for associating a metadata value to the described resource, folksonomies make no requirement in that sense, and the justification for a specific term on a given resource must be extracted *a posteriori* by examining and making hypothesis about the tag and the context in which it was specified (e.g. the other tags on the same resource, the tags used by others for the same resource, etc.).

A popular approach is to try to map folksonomic tags into a restricted number of categories, be they the sixteen of Dublin Core [8], seven [9], five [10], three [7] or just two [11]. But whatever the number, the attribution of a tag to a category is a complex and subjective issue requiring massive manual work, a substantial normalisation of the variants [8], some careful consideration of the context of the tagging and often a human visiting and analysing the described resource.

In this paper we present a humbler vision, but yet a bolder research proposition: even if we reduce ourselves to just two main categories (and even weaker than Strohmaier, Korner and Kern's ones [11]), and if we acknowledge that not even humans would reach an acceptable agreement on the interpretation of many tags, is it possible to create a computational approach that assigns tags to one or the other category with some accuracy and trustworthiness? Is it possible, computationally, to distinguish tags between *topical* (i.e., talking about the topic of the resource) and *non-topical* (i.e., describing aspects of the resource that are not related to its topic)?

In order to answer to this question, we collected a large sample of folksonomic tags from Delicious (21M tags, for a total of 1.2M unique terms). The folksonomy of Delicious is interesting to study because of the open nature of the Web documents that are tagged by the users. Since the website does not enforce any particular constraint, such documents can be regular HTML pages, but also videos, audio files, interactive games etc. Moreover, the users of Delicious tag documents coming from all kinds of sources, as opposed to other folksonomic aggregators where users typically tag their own self-made content (e.g., photos in Flickr and drawings in Deviantart).

We developed an algorithm, called *Non-Topicality by Distributional Semantics* (*NTDS*), that classifies the collected tags using a pure statistical approach. We validated its outcomes against a reference classification on a limited number of terms in three separate tests. In all three cases, the reference classification was provided by humans classifying the tags manually and according to an intuitive understanding of topicality (test 1), to a more precise and algorithmic definition of topicality (test 2) and to the expectation that the primary attribution of the tags by their authors was topical or non-topical (test 3). The analysis of the outcomes of the NTDS algorithm shows, in some cases, a consistent disagreement between humans about what constitutes a topical tag, and suggests the rise of a new category of overly generic tags that we named *umbrella tags*. Such kind of tags is the actual source of disagreement between humans and, thus, between humans and the NTDS algorithm.

The rest of the paper is organised as follows. In Section 2 we present recent works related to tag classification and the emerging of ontologies from social networks. In Section 3 we introduce our classification for tags, while in Section 4 we describe how we collected the data to test and propose NTDS for the identification of *topical* and *non-topical* tags based on natural language processing (NLP) tools. In Section 5 we introduce the tests and we discuss the outcome of NTDS in all aforementioned three cases. Finally, in Section 6 we conclude the paper sketching out some future developments of our work.

2. Related works

There exist several works that suggest different kinds of classifications according to behavioural and social aspects involved when users tag some resource.

Some of them try to use existing categories, the most obvious choice being the Dublin Core. In [8], for instance, the result of a pilot study [12] is contextualized, in which 311 tags (for 1141 occurrences in total) were mapped into the 16 terms of Dublin Core [13], with complex results: the majority of the tags (90,5%) ended up as *dc:subject*, 14 of the 16 elements ended up with at least one value, and several tags could not find a way into the element set, so that several new elements are proposed, such as Action (e.g. *toread*), Rate (e.g. *good*), Depth (e.g. *overview*) and Usage (e.g. *class*).

Among those creating their own categories, [9] provides a classification of tags into seven different sets:

- identifying what (or who) the resource is about: the topics of bookmarked items;
- identifying what the resource is, what kind of thing a bookmarked item is, such as article, or blog;
- identifying who owns the resource;
- refining categories: numbers and quantities;
- identifying qualities or characteristics, such as *funny* or *inspirational*;
- self reference, such as *mystuff* or *mycomments*;
- task organizing, such as *toread* and *jobsearch*.

Starting from these seven categories, Sen *et al.* derive three different categories that describe *factual* (that identifies facts about, for instance, books such as characters, places, etc.), *subjective* (that express user opinions related to something, for instance a book) and *personal* characteristics (that describe the intended audience or feeling of users who applied the tag) of tags [14].

Xu *et al.* propose another classification based on five different categories [10]:

- *content-based*, that describes the content of an item;
- *context-based*, that describes the context of an item in which it was created or saved, e.g. locations;
- *attribute*, that introduce inherent characteristics of an item without being part of the content, e.g. the source of an article;
- *subjective*, i.e., that expresses users' opinions;
- *organisational*, identifying personal stuff or that reminds certain task.

Gupta *et al.* [15] expand Xu *et al.*'s classification by adding six additional categories of tags:

- *ownership*, tags that specify the owner of the resource;
- *purpose*, which denote specific functions that do not relate with the content of a resource and refers to information seeking tasks of users;
- *factual*, that identify facts about people or objects;
- *personal*, specifying an intended audience that relates with tag appliers themselves;
- *self-referential*, i.e., tags to resources that refer to themselves;
- *tag bundles*, that refers to tags that are applied to other tags, in order to create a hierarchical organisation of folksonomies.

In [7], on the other hand, a model of three categories is used:

- *subject tags*, bearing some evidences of the development of a reasonable consensus on the *aboutness* [16] of the studied resources;
- *affective tags*, describing an emotional state; and
- *time, task or project related tags*, e.g., compound words such as *toread* and *todo* and appearing to indicate a desire to combine information about tasks and activities with subject classification terms.

The authors also noted and decided to ignore a small set of tags consisting of prepositions, conjunctions and other parts of speech from tag phrases which were separated by the system into individual tags.

Another approach is to identify the category of the tag by studying the intention of the human tagger. In [11] a distinction in the underlying purpose of tagging is made between categorisers and describers, the first being inclined to frequently reuse elements from a limited vocabulary towards an eventual browsing of few well-organized sets of resources, while the latter fostering a rich spread of often similar terms that are meant to be later retrieved via unforeseeable search items, with unplanned reuse of terms, little restriction in the vocabulary but on the contrary a rich selection of alternative terms with the same meaning.

Similarly, Yang *et al.* analysed Twitter *hashtags*, tags that are used to describe or characterise a tweet (e.g., *#iphone*) [17]. They provide a macroscopic analysis of two alternative roles that hashtags may have in Twitter, i.e., being *content indicators* or *community membership indicators*. For instance, the hashtags *#1Million600K* and *#RenewUI* in President Barack Obama's tweet "There are currently #1Million600K American job seekers without unemployment insurance. This must end now: http://ofa.bo/p0D #RenewUI"[1] are a content indicator and a community membership indicator respectively.

In Table 1, we summarise all the classifications introduced in this section, which represent the main research achievements in this field.

Table 1. Previous works on tag classifications and related tag categories.

Work	Categories
Golder and Huberman [9]	• Aboutness • Type • Ownership • Number/quantities • Qualities/characteristics • Self-referential • Task (a.k.a. Organisational)
Sen *et al.* [14]	• Factual • Subjective • Personal
Xu *et al.* [10]	• Content-based (a.k.a. Aboutness) • Context-based • Attribute • Subjective • Organisational
Gupta *et al.* [15]	All the categories mentioned in [10] plus the following ones: • Ownership • Purpose • Factual • Personal • Self-referential • Tag bundles
Kipp [7]	• Subject (a.k.a. Aboutness) • Affective • Time/task/project related (a.k.a. Organisational)
Strohmaier *et al.* [11]	This classification applies to human taggers rather than to tags: • Categorisers • Describers
Yang *et al.* [17]	• Content indicator (a.k.a. Aboutness) • Community membership indicator

3. About the (non-)topicality of tags

In this section we give a proper definition of *topical* and *non-topical* tags and provide some insights on their possible applications in the context of the Semantic Web and Ontology Engineering.

3.1. A simple model

The classification models introduced in Section 2 are wildly different in their structure, aim and content, but they seem in agreement on one aspect: even with different names (i.e., *subject*, *dc:subject*, *"what the resource is about"*, *content-*

based tags or *descriptive* tags) this category appears in all models, and appears to subsume the idea of *aboutness* [16], i.e., what the document is about, one of the most important concepts in information science. While we may all agree on an informal, approximate idea of the aboutness of a document, of course the definition is subtle and is not completely overlapping in at least some of the categorisations above described. To avoid confusion, we will call this category with yet another term, *topicality*, also used in information science although with different nuances, and will call *topical tag* any term that associates the resource to a topic (e.g., a knowledge domain) that is appropriate to its content.

Even more diverse is the list of the other categories, which span over intentionality, subjectivity, quality and context of the tagging. There is little or no agreement over these categories and most probably no agreement is actually possible, so we will not attempt to impose one, and will refer to these tags with a negative term, as *non-topical tags*, to refer to terms that are ***not*** identifying a knowledge domain appropriate for the content of the resource. This is a very simple taxonomy of two elements only, even narrower than the one in [11], since categorizing tags (i.e., such tags associated to a resource by a categoriser) do include, to a certain extent, even topical terms, although of a very general type.

Yet we can assert some intuitive characteristics of topical vs. non-topical tags:

- *topical tags* (i.e., tags that **do** talk about the actual *content* of a document) describe potentially any domain of the human knowledge, and may use a large portion of the language to do so; furthermore, topical tags over the same resource are often semantically related, as they represent glimpses over a conceptualization of the specific domain being talked about in the document. On the contrary,
- *non-topical tags* (i.e., tags that **do not** talk about the *content* of a document) are probably coming from a *smaller*, albeit vast, portion of the language, describing a more restricted set of issues connected to opinions, audiences, sources, tasks and context than topical ones; yet they are not associated to a specific domain of the human knowledge and in fact can be found listed among the tags of an extremely wide range of resources.

Of course, variability exists in how tags are used: while some tags (e.g., *chocolate*) are most evidently of a topical nature, and others (e.g., *toread*) are hardly imaginable as topics, many can (and in fact do) position themselves in an intermediate position, being used sometimes as topics (e.g. *kids* in the context of an article *about* young humans) and sometimes not (e.g. *kids* in the context of a web page describing a toy or an amusement park *for* young humans).

As introduced in Section 1, our intuition is that it could be interesting to provide an automatic approach to identify with an acceptable confidence the *non-topical* tags, i.e., those that are commonly used with the intent of asserting facts *not* about the content of the resource. Intuitively, non-topical tags will be present identically in many different domains, and therefore have patterns of contiguity completely different from a topical tag.

3.2. Towards the creation of lightweight non-topical ontologies

The identification of non-topical tags as deriving from the use that tag authors made of them, rather than from intrinsic qualities of the tag, suggests the existence of emerging ontologies implicitly used by the community. In the domain of Semantic Web research, an important step forward about folksonomies and emergent ontologies was proposed by Mika [18]. In his work, Mika proposes a tripartite model for the analysis of ontologies so as to take into account the social dimension, besides the description of concepts and instances. In addition, he shows how it is possible to identify emergent ontologies from folksonomies with emergent semantics approaches.

While on the one hand the tags grouped according to their co-occurrence on items, as suggested by Mika [18] and Specia and Motta [19], show emerging clusters that seem to describe topical characteristics of the tag space, thus implicitly define sort of domain ontologies, on the other hand the identification of non-topical tags seems to let a different kind of lightweight ontologies emerge.

For instance, the tag *kids* – when considered as a non-topical tag – may imply the existence of a particular class describing the intended audience for the related items, while the tag *newyorktimes* may refer to the particular source from which the described resource is derived. Lightweight non-topical ontologies (i.e., ontologies that do not talk about topical attributes of resources) emerge from the identification of non-topical tags according to a number of particular purposes, and can be organized according to the non-topical characteristics of many of the classifications described in Section 2 (including the one described in [20]) or other subject headings (e.g., along the line of [21]), without the need to choose one ontology against the other.

This open-ended approach suggests us also to evaluate non-topicality as either time-independent or dependent (i.e., *rigid or anti-rigid*, according to the OntoClean methodology [22]), since users can decide to apply them permanently or only temporarily[2]. For instance, the tag *toread* (usually referring to a document that should be read later on) may subsequently (e.g., after the document has in fact been read) become pointless or wrong. In these cases, ontology

evolution [23] approaches such as [24] may be applied to folksonomies so as to study how emerging non-topical ontologies evolve during time.

4. An algorithm for identifying (non-)topical tags

In this section we introduce our algorithm (i.e., *Non-Topical by Distributional Semantics*, a.k.a. *NTDS*) and the experimental setting through which we assess the quality of NTDS outcomes.

4.1. Collecting tags

In order to follow through with our intent to design and test an algorithm that can determine the topicality or non-topicality of folksonomic tags, we collected a portion of a folksonomy using tags extracted from Delicious.

A script was run to read Delicious news feeds between August 2010 and September 2010. The script ran for about six weeks, gathering information about slightly less that 1.3 million documents. The first pass of the crawling process collected a sequence of tagging events of the type "user assigns label to URL", independently from their topic. Note that this process is different from a keyword-based search, but rather it is the systematic capture of the continuous flux of activity on Delicious. The news feeds contain only data about the first time a user tags a document, so as a second step we subsequently took every document described in the first batch and we queried Delicious about it accessing all its tags along with the user name of the tagger and the timestamps of the tagging. No manual selection or post-processing has been performed on the tag set. This process took about a week using ten different machines simultaneously.

Overall, we gathered information about 1,280,686 documents, for a total of 21,408,652 tags (16.7 tags per document on average), 1,205,958 unique tags, 491,702 users and 7,034,524 tagging events[3]. The distribution of the number of documents per number of tags is shown (in a logarithmic scale) in Figure 1. The dataset was stored as an XML file (950MB) according to the following format:

```
<tags t="1229008773" u="[username]" href="005cb474bfc10f41036b543f042ae791">
  <t>jquery</t>
  <t>webdesign</t>
  <t>navigation</t>
</tags>
```

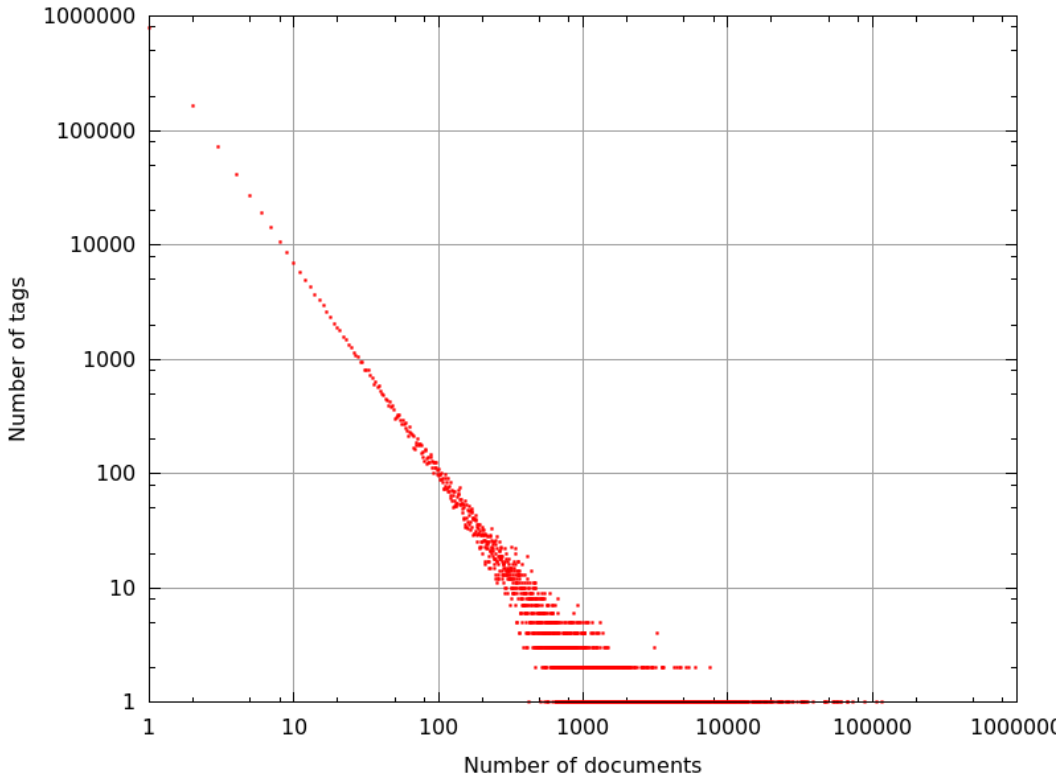

Figure 1. The distribution of documents per tag in our dataset. It shows the number of tags in our collection (axis y) associated to exactly x documents.

Every element *tags* represents a tagging event, i.e., the activity of a *user* who tagged a *document* with zero or more *tags* (specified through the element *t*). The three attributes of the element *tags* are:

- *t*, i.e., the timestamp of the tagging event in UNIX time format;
- *u*, i.e., the username of the tagger in Delicious;
- *href*, i.e., the md5 hash of the document URL that was tagged.

The procedure described here aims at collecting a large dataset with high coverage over topics, i.e., containing tags relative to as many domains as possible. As such, should the need ever arise to investigate folksonomic datasets in a different manner, for example looking for domain-specific or language specific ones subsets, our crawling algorithm will have to be adapted.

4.2. NTDS: Non-Topicality by Distributional Semantics

We base our reasoning on the idea that the non-topical tags of a given folksonomy can emerge from the analysis of the topological organisation of the folksonomy itself. Roughly speaking, the idea is to organise the tag space in clusters where tags are grouped according to how many times they are used together to annotate resources, which then become connected through *hubs*, i.e., tags that connect several clusters between them. Our intuition is that the hubs represent non-topical tags of the folksonomy, since they connect many different contexts. The actual approach we used is introduced as follows.

Our method is grounded on the Latent Semantic Analysis (LSA) firstly introduced by Deerwester *et al.* [25] and largely used by the NLP and information retrieval communities. The occurrence of each word is represented as a vector of words co-occurring with it in the same context (different contexts can be used, e.g. some neighbouring words, a sentence or even an entire document). All vectors are thus collected in a *[word X word in context]* matrix *A*. This is a

huge and sparse matrix, because each original vector spans through the entire lexicon of the target language. Applying a Singular Value Decomposition technique to *A* we map each vector into a subspace, called the *word-space*, with reduced dimensionality *k*, keeping most of the original distributional information and making evident high order latent relationships between words. A good value for *k* (=300) has been empirically determined in the LSA literature to allow for the extraction of latent high-order relationships between words.

Each word is then mapped into a k-dimensional vector in order to efficiently compare it to other word-vectors to find similarities using the cosine distance:

$$Sim(w_1, w_2) = \frac{w_1 * w_2}{\|w_1\| \cdot \|w_2\|} \quad (1)$$

where *w1*w2* is the Euclidean scalar product and ||*w*|| is the norm of the vector *w*. The closer the vectors in the k-dimensional space are, the more the distributional behaviours of the two words are similar. So, appropriate clustering methods can identify sets of similar words.

Various other word-space models can be used instead of LSA, for example *Hyperspace Analogue to Language* (HAL) [26] and *Latent Dirichlet Allocation* (LDA) [27]. The interested reader can have a look at the review on word-spaces from [28]. We chose LSA for our experiments mainly for its simplicity and because of the high availability of this model. The global results, as well as the conclusions, should not be heavily affected by a different choice of the word-space model.

Starting from this standard technique, Widdows [29] studied the properties of this word-spaces, and the networks of words generated by connecting each word to its most similar words.

We applied this idea by building a word-space considering the content of the tagging events (i.e., all the tags associated by a single user to a specific document) as if it constituted a "text-sentence" in the standard word-space models. Thus the context for building the distributional model became this set of tags. Then, for each tag *T* we extracted the 200 most similar tags through the word-space model. Each tag *T'* in the set of immediate neighbours, *NT*, is connected with *T* with an edge weighted using the similarity between the two tags *Sim(T,T')*. We then iterated the same process considering each tag *T'* in *NT* enriching the graph *GT* with many more edges.

Given the network of tag *GT* we can observe, in line with the considerations of, among many, Widdows [29] and Heyer *et al*. [30], that tags used only in few specific contexts tend to be placed in a highly connected subgraph, while tags used in different contexts tend to be hubs between different highly connected subgraphs. In other words, GT exhibits the property to be a *small-world* graph in the sense defined by Watts and Strogatz [31].

Widdows [29] and Dorow *et al.* [29] introduce a measure for unweighted graphs, the node *curvature* (also called *clustering coefficient* [31]), to quantify the tags' behaviour and measure the cohesiveness of the tags' neighbourhoods The curvature of a node (tag, in our case) *t*, *NC(t)*, is defined by:

$$NC(t) = \frac{\#(triangles\ t\ participates\ in)}{\#(triangles\ t\ could\ participates\ in)} = \frac{2 \cdot \sum_{Triangle(t,ta,tb)} 1}{nn \cdot (nn-1)} \quad (2)$$

where *Triangle(ta,tb,tc)* marks a triangle between the nodes *ta*, *tb* and *tc*. The node curvature assumes values between 0 and 1. A value of 0 occurs if there is no link between any of the node's neighbours, and a node has a curvature of 1 if all its neighbours are linked (see Figure 2 for some artificial examples). Thus a node that exhibits a high curvature is part of a strongly interconnected subnet, i.e., a small world, while a node exhibiting small curvature should be a hub between different subnets (i.e., different small worlds).

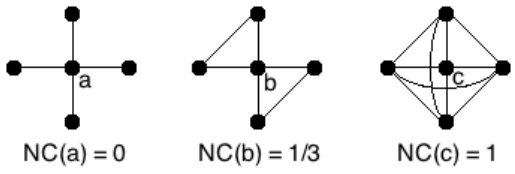

Figure 2. Different node curvature, as a function of neighbouring nodes, as defined in [29].

JIS Accepted for Publication
By the Journal of Information Science: http://jis.sagepub.co.uk

In *Non-Topicality by Distributional Semantics* (*NTDS*), we extended this measure for weighted graphs as follows: instead of counting 1 for each triangle to which *t* takes part, we add to the sum the *minimum edge weight* among the edges that compose the considered triangle. Thus, if the node *t* participates in a triangle with nodes *ta* and *tb* the *Weighted Node Curvature* (*WNC*) is defined as:

$$WNC(t) = \frac{2 \cdot \sum_{Triangle(t,ta,tb)} \min(w_{t-ta}, w_{t-tb}, w_{ta-tb})}{nn \cdot (nn-1)} \quad (3)$$

where $w_{t-t'}$ is the weight between node *t* and *t'* and *nn* is the number of neighbours of *t.* Extending the computation of node curvature considering weighted edges representing node similarity results in a more precise definition of node/neighbourhoods cohesiveness and allows for a better small-world identification.

In our hypothesis, if the WNC of a tag is small the tag connects different domains (i.e., it shows a non-topical characterization), as shown in Figure 3, while WNC scores closer to 1 should be hints of topical tags.

Once we have computed the word-space model for all our documents, a rather slow process, the proposed algorithm can compute WNC quite quickly, processing a large bunch of tags in few seconds.

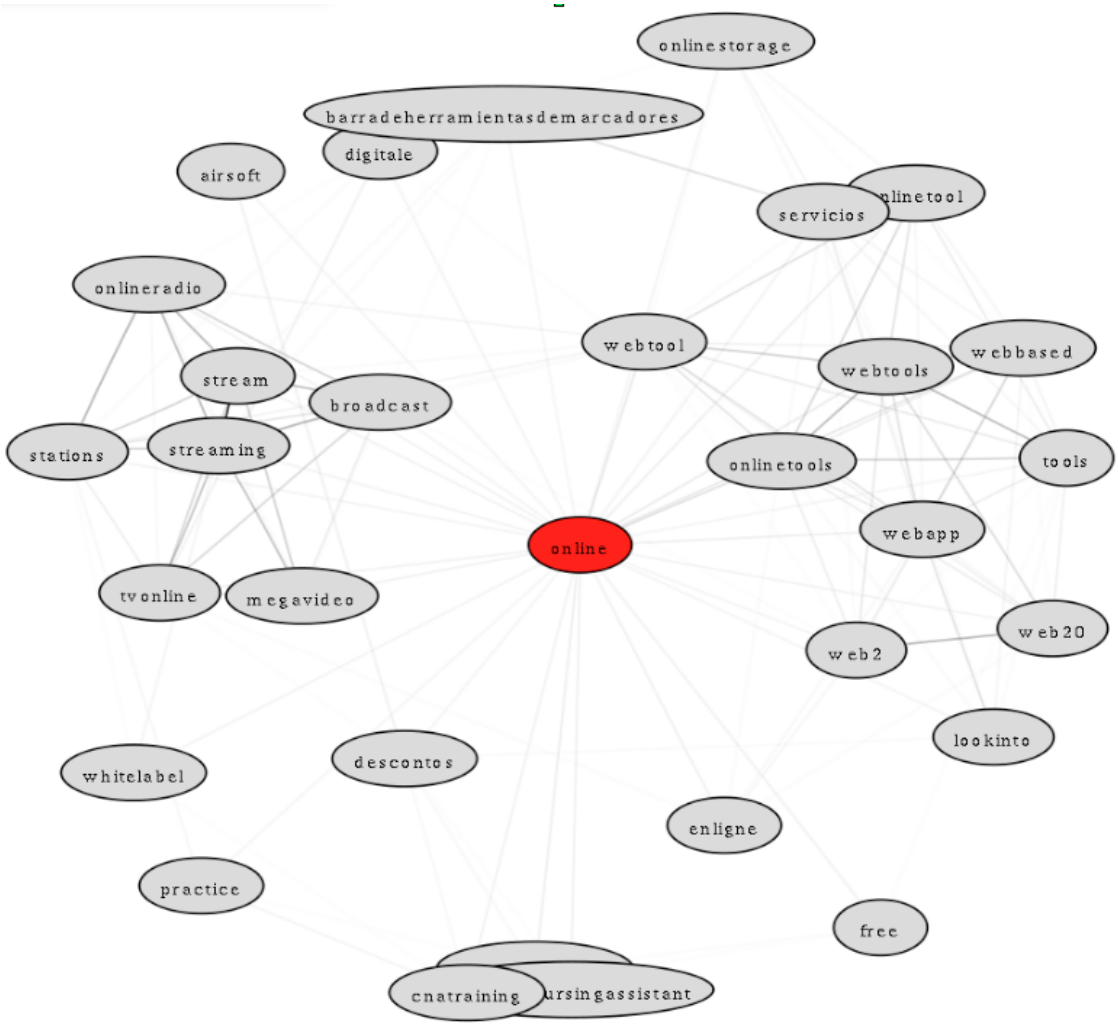

Figure 3. A snapshot of the tag graph produced by NTDS, where the tag "online" has been recognised as non-topical tag since it links several tag clusters.

5. Evaluation of NTDS

In order to evaluate the qualities of NTDS, we made three separate and independent evaluations on the quality of the output of the NTDS algorithm[4].

In the first test, the output of NTDS was used against comparable evaluations based on a small set of randomly chosen tags from the same input data performed by a limited number of human cataloguers. In this test, a vaguely defined common-sense definition of topical tags was used by the human cataloguers, who had access to the full set of information available on the tag, including the resources themselves which the tags where associated to. In Section 5.1 we discuss in detail the output of such test, which returned interesting patterns of similarities between humans and algorithm, but provided evident and undeniable issues in the inter-human agreements that we attributed to an excessively vague definition of topicality.

In the second test we provided human cataloguers with a more precise, almost algorithmic definition of topicality, and applied it to a different, and explicitly chosen, set of tags. This improved the inter-human agreement (in as much as they strictly followed the proposed algorithm – although it created dissatisfactions in a few cases), but created clear outliers in the comparison with the output of the NTDS algorithm. In Section 5.2 we discuss the details of this test.

We then decided to add a folksonomic flavour to our test, and verify the predictive value of the output of the NTDS algorithm, i.e., how well its values agree with the expected characterisation of a tag by humans, who most often try to evaluate the nature of a tag without accessing the resources that would disambiguate it. A larger number of tags at the extremes of our ranges were evaluated in a simpler fashion, as discussed in detail in Section 5.3.

5.1. The first test

In a first test we compared the output of the NTDS algorithm against a reference evaluation provided by three human cataloguers. Three individuals, with some experience in Delicious and a computer science background, were asked to evaluate the quality of the output of the NTDS algorithm by examining a random selection of 10 of the 1000 most frequently used elements of our collection of 1.2M tags. For each of them, we selected 20 documents associated to the tag and asked the cataloguers to evaluate whether the tag was or was not topical for the document. We gave no strict definition of topicality, and relied instead on the common interpretation of the term as presented in Section 3.1.

The tags considered in the experiment were *books*, *environment*, *free*, *game*, *healthcare*, *howto*, *online*, *philosophy*, *python* and *vegetarian*. The WNC scores returned by the NTDS algorithm are shown in Table 2. As specified, a low score indicates a non-topical characterization of the tag, and a higher score, on the contrary, a characterization of the tag as topical.

Table 2. WNC scores returned by the NTDS algorithm for the ten tags used in the first experiment.

Tag	WNC score
books	0.32
environment	0.42
free	0.09
game	0.40
healthcare	0.59
howto	0.09
online	0.05
philosophy	0.56
python	0.41
vegetarian	0.72

We then asked cataloguers to classify each of the twenty occurrences of a particular tag (one per document) as either topical or non-topical. Given these classifications, we calculated two measures, *score* and *match*, defined as follows:

$$score(u, tag) = \frac{|\{doc_k \mid topical(tag, doc_k, u) = true\}|}{|\{doc_k \mid exists(tag, doc_k) = true\}|} \quad (4)$$

$$match(u_1, u_2, tag) = \frac{|\{doc_k \mid topical(tag,doc_k,u_1) = topical(tag,doc_k,u_2)\}|}{|\{doc_k \mid exists(tag,doc_k) = true\}|} \quad (5)$$

The supporting function *topical(tag,doc,user)* returns true if *tag* was classified as topical in document *doc* by *user*, and *exists(tag,doc)* returns true if *doc* was annotated with *tag*. The score measure is the mean of the annotations a user *u* made on *tag* – here, too, lower values implies non-topicality, higher values implies more topicality. Finally, the match measure describes how similarly users *u1* and *u2* categorised *tag* – producing higher values when they categorised it in a similar way, and lower values otherwise. The score and match values are shown in Table 3 and Table 4 respectively.

Table 3. The scores resulting from cataloguers' annotations of the first experiment. Lower values implies non-topicality, higher values implies more topicality.

Tag	cataloguer1	cataloguer2	cataloguer3
books	0.32	0.88	0.66
environment	0.34	0.62	0.98
free	0.00	0.02	0.04
game	0.86	0.62	0.76
healthcare	0.68	0.52	0.98
howto	0.00	0.12	0.08
online	0.02	0.00	0.08
philosophy	0.64	0.52	0.96
python	0.58	0.38	0.74
vegetarian	0.20	0.08	1.00

Table 4. The matches among cataloguers in the first experiment. Higher values implies that the two users categorised a tag in a similar way, and lower values otherwise.

Tag	cataloguer1 vs. cataloguer3	cataloguer1 vs. cataloguer2	cataloguer2 vs. cataloguer3
books	0.58	0.44	0.66
environment	0.32	0.60	0.60
free	0.96	0.98	0.94
game	0.62	0.68	0.78
healthcare	0.70	0.72	0.54
howto	0.92	0.88	0.92
online	0.90	0.98	0.92
philosophy	0.68	0.68	0.52
python	0.76	0.68	0.52
vegetarian	0.20	0.84	0.08

Our expectation was that there would be similarities between the topicality as expressed by the human score and by the WNC score returned by NTDS. In Figure 4 we show the results of the first experiment. All tags are presented so that their *X* coordinate represents the WNC score and the *Y* coordinate represents the value of the mean of the human scores. Additionally, the size of the circle represents the value variability between humans – the smaller the circle, the greater the agreement – calculated as the value of the mean of the human matches for a particular tag[5].

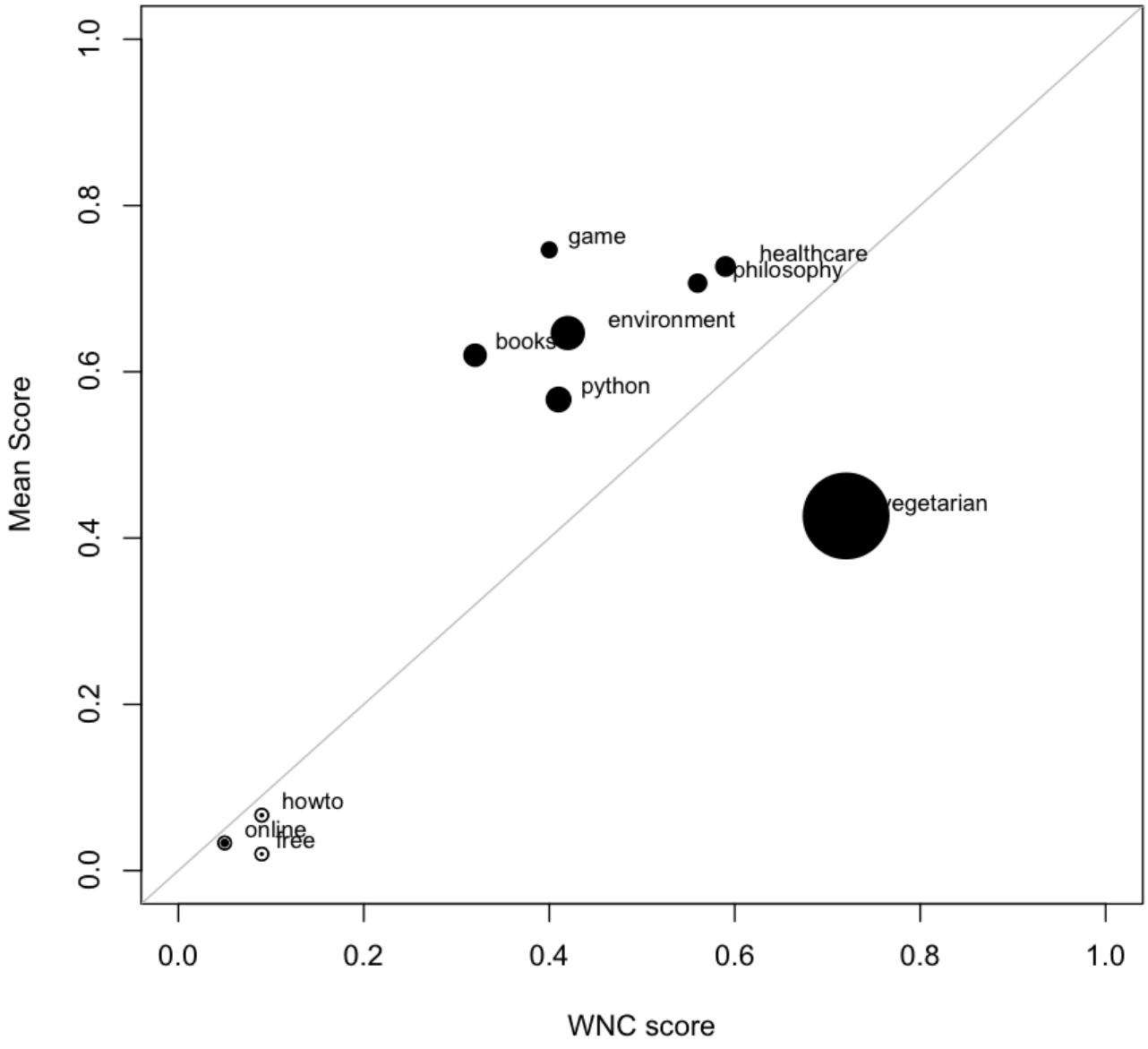

Figure 4. The tags of the first test ordered by WNC score (X axis) and the mean of human score values (Y axis). The size of the circles marks the spread in agreement (i.e., the mean of human match values) between cataloguers (the larger the area, the smaller the agreement).

Evaluating the result of the test, we noticed that the algorithm performed well within the boundaries of the differences between humans, and in a few cases (i.e., *howto*, *free*, *online*, *philosophy* and *healthcare*) it was exactly overlapping some of their opinions. However, while the human agreement for tags with lower scores was very strong (i.e., the bottom-left cluster), there still were some differences in opinion between humans for those tags perceived as more topical (the top-right cluster), which made the results hard to interpret.

Wrap-up discussions and interviews showed substantial disagreements between the human cataloguers about what constituted a topical tag: for instance, no agreement was reached about whether "game" is a topical tag for a web page that ***is*** a game, but ***does not talk about*** this one game or games in general (e.g., a web page containing a Flash online game and nothing else). We decided therefore that the common interpretation of "topical" could be too imprecise or not sufficiently shared, and decided to run a second test with a stricter definition of what constitutes a topical tag.

Furthermore, the random selection did not create a balanced distribution of tags according to their WNC scores and we decided to look for a more evenly distributed selection by choosing test tags explicitly.

5.2. The second test

The second test was performed again against a reference evaluation provided by three human cataloguers. This time 20 of the 1000 most common tags (except the 10 tags already used in the first experiment) were selected so as to provide a reasonable distribution in their WNC ranking. Then 50 documents were selected for each of them, and we asked three cataloguers to provide their own reference evaluation of the tags based on the selected documents. The effort for such test is heavy, as it requires 1000 web pages to be accessed and, at the very least, cursorily scanned to determine the justification for the use of the corresponding tag, and so average run time for each of our testers was over 18 hours, which explains the actual number of humans we were able to enrol.

In order to help users to better discriminate the use of the tags, we provided them with a precise guideline in the form of a human-executable algorithm.

Given a document D and one of its tags X, we say that X is a topical tag in D if and only if:

- *X answers the question "Does D talks about X?"; or*
- *X answers the question "Does D talks about Y, where Y is a particular instance of X?"; or*
- *If Z is an hyponym of X (by sense or according to a thesaurus such as Wordnet [33]) and Z is a topical tag in D according to one of the above rules then X itself is a topical tag in D.*

If X was not identified as topical according to any of the previous rules, then X was a non-topical tag in D.

For instance, on a page with a discussion about some powerups of *Call Of Duty* (a first person shooter of some fame), all of the following would be considered topical tags: *callofduty*, *powerup* (for rule 1), *FPS* (for rule 2), *game* (for rule 3)[6]. Of course these guidelines are still vague enough to allow for subjective interpretation of a tag. In particular, the meaning of "talks about" was left to users, as well as the handling of rule 3 for terms that are not actually present in Wordnet (e.g., "webdev").

The list of the tags considered in the experiment was: *architecture*, *awesome*, *bible*, *blog*, *business*, *chocolate*, *copyright*, *gardening*, *geometry*, *guitar*, *inspiration*, *iphone*, *linkedin*, *logo*, *resource*, *science*, *university*, *upload*, *webcomic* and *webdev*. The WNC scores returned by the NTDS algorithm are shown in Table 5.

Table 5. WNC scores returned by the NTDS algorithm for the twenty tags used in the second experiment.

Tag	WNC score
architecture	0.45
awesome	0.14
bible	0.69
blog	0.13
business	0.21
chocolate	0.80
copyright	0.47
gardening	0.62
geometry	0.53
guitar	0.57
inspiration	0.19
iphone	0.28
linkedin	0.13
logo	0.31
resource	0.07
science	0.19
university	0.53
upload	0.28
webcomic	0.48
webdev	0.09

As before, we asked cataloguers to classify each of the fifty occurrences (one per document) of a particular tag as either topical or non-topical and we then calculated again the related *score* and *match* values, shown in Table 6 and Table 7 respectively.

Table 6. The scores resulting from cataloguers' annotations of the second experiment. Lower values implies non-topicality, higher values implies more topicality.

Tag	cataloguer1	cataloguer2	cataloguer3
architecture	0.60	0.64	0.90
awesome	0.00	0.02	0.02
bible	0.42	0.56	0.66
blog	0.08	0.08	0.08
business	0.40	0.66	0.52
chocolate	0.26	0.18	0.94
copyright	0.74	0.76	0.86
gardening	0.68	0.50	0.94
geometry	0.40	0.86	0.88
guitar	0.40	0.48	0.68
inspiration	0.02	0.00	0.02
iphone	0.12	0.24	0.80
linkedin	0.66	0.68	0.62
logo	0.38	0.70	0.56
resource	0.00	0.50	0.18
science	0.56	0.52	0.78
university	0.54	0.28	0.36
upload	0.52	0.74	0.44
webcomic	0.02	0.20	0.04
webdev	0.80	0.66	0.76

Table 7. The matches among cataloguers in the second experiment. Higher values implies that the two users categorised a tag in a similar way, and lower values otherwise.

Tag	cataloguer1 vs. cataloguer3	cataloguer1 vs. cataloguer2	cataloguer2 vs. cataloguer3
architecture	0.66	0.60	0.62
awesome	0.98	0.96	0.94
bible	0.58	0.82	0.64
blog	0.92	0.94	0.94
business	0.50	0.64	0.52
chocolate	0.28	0.72	0.18
copyright	0.88	0.82	0.78
gardening	0.96	0.98	0.94
geometry	0.74	0.66	0.52
guitar	0.48	0.54	0.82
inspiration	0.72	0.80	0.72
iphone	0.92	0.88	0.92
linkedin	0.96	0.98	0.98
logo	0.32	0.84	0.44
resource	0.82	0.90	0.84
science	0.64	0.60	0.76
university	0.90	0.98	0.92
upload	0.82	0.46	0.52
webcomic	0.64	0.56	0.48
webdev	0.62	0.72	0.74

We noticed that the algorithm did perform worse than the previous test when comparing humans' scores. Contrarily to what we expected, the new guidelines did not noticeably increase the inter-human agreement, and rather created a strange situation in the comparison against the WNC scores by the NTSD algorithm. As shown in Figure 5, in fact, we see three main clusters of tags: the bottom left and top right ones represent a good similarity between the cataloguers and NTDS, while the top left one shows a rather radical difference, where humans classified the terms as more or less topical while the algorithm classified them as clearly non-topical. A further analysis and an interview with the cataloguers helped to shed light on the occurrence: these are all terms that in many cases the cataloguers considered as topical because of a literal interpretation of rule 3, and that were felt as very general and unspecific for the document[7]: for instance, the homepage of the Queensland museum[8] that was interpreted as "science" or a page about Adobe Creative Suite 5.5[9] that was interpreted as "webdev".

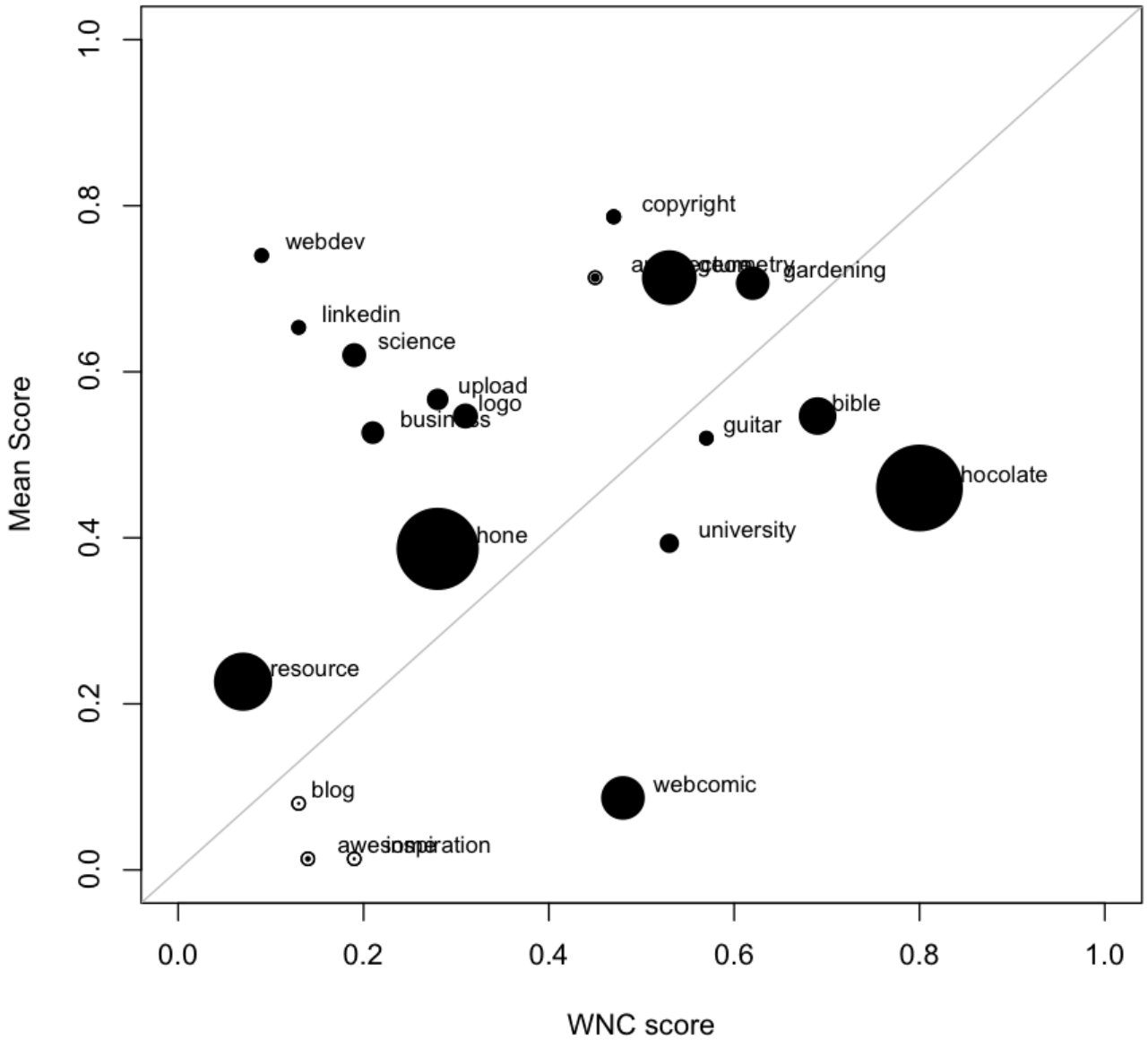

Figure 5. The tags of the second test ordered by WNC score (X axis) and the mean of human scores (Y axis). The size of the circles marks the spread in agreement (i.e., the mean of human matches) between cataloguers (the larger the area, the smaller the agreement).

These terms were interpreted topical in the classification of the human testers because of the excessively generic rule 3, and were considered non-topical in the NSTD algorithm – probably because they were used as introductory terms in many different specialised contexts, so that they became hubs of separate clusters because of their very generality. We call these terms *umbrella tags*, and we postulate that in traditional classification science the problem has rarely if ever occurred because of the constant thrive of its practitioners to use the most specific term available in the thesaurus, which most often constituted a leaf in the tree of the available terms. Distinguishing umbrella tags from non-topical tags becomes therefore an open topic of discussion within our research framework. In addition, as a side note, in principle the algorithmic approach used in NTDS could also confuse topical tags that are very general, such as *science*, with *unpopular* non-topical tags. In particular, unpopular non-topical tags may not connect several clusters due to their rare

usage, and they may be potentially confused with general tags connecting a similar number of clusters. A further analysis in this direction should be addressed in future studies as well.

5.3. The third experiment

Testing the agreement between humans and NTDS in the classification of a tag is a way to test, basically, whether the algorithm can guess the opinion of the author of the tagging about the tags themselves. This task has proven to be long, fatiguing, and for practical reasons only applicable to a rather limited set of tags.

Yet, one of the fundamental assumptions of folksonomies can help us in designing another test on a larger set of tags that does not involve a comparable amount of work as the previous ones. It has been demonstrated (e.g., in [32]) that sufficiently large folksonomies tend to become stable, i.e., a coherent categorisation scheme emerges from collaborative tagging so that tags organise themselves in a power law that shows a statistically reliable agreement in the evaluation of the resource, in particular between taggers and searchers. In practice this means that searchers' expectation of the meaning of a tag is a reliable indicator of the intended meaning of the taggers.

Therefore, for the third test we decided to test the quality of the output of the NTDS algorithm against the expectation of topicality or non-topicality that searchers have when shown tags that were used to describe documents, instead of considering the original characterisation of the tags in the context of specific documents. Since this requires no access to the documents – i.e., only a first-glance evaluation of the meaning and characteristics of the tags is actually required – a quick and low-impact test could be conducted.

We selected 60 different tags within the 1000 most used tags (excluding those already used in the previous tests), 20 in the group with the lowest score (WNC < 0.2, i.e., putatively, clearly non-topical), 20 in the group with the highest score (WNC > 0.8, i.e., putatively, clearly topical) and 20 close to the median score (WNC between 0.4 and 0.6, i.e., putatively, intrinsically ambiguous). The WNC scores returned by the NTDS algorithm are shown in Table 8.

Table 8. WNC scores returned by the NTDS algorithm for the sixty tags used in the third experiment.

Tag	WNC score
art	0.28
arthureames	0.89
arthurmerlin	0.90
aspnet	0.50
australia	0.52
blogs	0.18
brendonryan	0.92
community	0.24
comparison	0.18
configuration	0.15
deancastiel	0.85
dessert	0.79
directory	0.16
editing	0.12
engine	0.17
examples	0.14
extensions	0.19
fandomsherlock	0.92
fashion	0.40
geo	0.40
graphic	0.39
graphics	0.19
harrydraco	0.93
harrypotter	0.89
hiking	0.55
hp	0.85
inception	0.89
internet	0.30

ipod	0.19
jaredjensen	0.87
javascript	0.18
kirkspock	0.91
kurtblaine	0.94
make	0.12
mckaysheppard	0.88
merlin	0.90
merlinarthur	0.90
new	0.17
pairingarthureames	0.89
portal	0.13
power	0.60
programming	0.11
read	0.22
safety	0.29
science	0.19
share	0.31
sherlock	0.92
sherlockjohn	0.92
social	0.19
socialmedia	0.36
ssl	0.36
startrek	0.91
stevedanny	0.82
support	0.13
teachers	0.30
twitter	0.40
urban	0.32
useful	0.06
viapackratius	0.13
website	0.20

After having sorted them in a random order, they were submitted to four potential searchers that were asked to evaluate how much they expected the tag to have been used by taggers as topical or non-topical according to the definitions provided in Section 3.1.

Values were disposed on a five-point Likert scale, with 1 representing "mostly or always non-topical", 2 being "more non-topical than topical", 3 being "similarly non-topical and topical", 4 being "more topical than non-topical", and 5 being "mostly or always topical". The results obtained from the users were then linearized between 0 and 1 (1 being 0.0, 2 being 0.25, 3 being 0.5, 4 being 0.75, and 5 being 1.0) – we assumed that the linearized values represent a reasonable approximation of the score assigned.

Contrarily to the previous experiments, here we used the following alternative implementation of the *score* and *match* functions to calculate directly the mean of these values for each tag:

$$score(Annots_{tag_k}) = \frac{\sum_{user_i \in Annots_{tag_k}} getValue(user_i, Annots_{tag_k})}{|\{user_i \mid user_i \in Annots_{tag_k}\}|} \quad (4)$$

$$match(Annots_{tag_k}) = \frac{|LikertCategories| - |getCategoriesUsedIn(Annots_{tag_k})|}{|LikertCategories| - 1} \quad (5)$$

The formula *score* takes as input the annotations made by humans for a particular tag tag_K and returns a score from 0 to 1 measuring the how much tag_K was annotated as non-topical (close to 0) or topical (close to 1). The formula *match* takes as input the annotations made by humans for tag_K and returns a score from 0 to 1 measuring the agreement between experts for that particular tag_K. *LikertCategories* is the set of the five categories used in the experiment (labelled from 1 to 5); the function *getValue* returns the converted the Likert category specified by a user for tag_K in the

appropriate 0-1 value (as shown in Table 9); while the function *getCategoriesUsedIn* returns the set of all Likert categories used by humans when annotating tag_K[10].

Table 9. The conversion of the cataloguers' annotations of the third experiment, given according to Likert categories, in the appropriate 0-1 values. Lower values implies non-topicality, higher values implies more topicality.

Tag	cataloguer1	cataloguer2	cataloguer3	cataloguer4
art	0.75	0.75	0.75	0.75
arthureames	1	0.75	1	1
arthurmerlin	1	1	1	1
aspnet	0.75	0.25	0.5	0.75
australia	1	0.5	1	0.5
blogs	0.25	0.5	0.25	0.25
brendonryan	1	1	1	1
community	0.25	0.75	0.5	0
comparison	0	0.75	0	0
configuration	0.75	0	0.25	0
deancastiel	1	1	1	1
dessert	1	1	1	0.5
directory	0	0.25	0.5	0.5
editing	1	0.25	0.75	0.25
engine	0.75	0.25	0.5	0.25
examples	0	0	0	0
extensions	0.75	0.25	0.25	0.25
fandomsherlock	1	1	1	1
fashion	1	0.5	0.75	0.75
geo	0.5	0.5	0.75	0.75
graphic	0.25	0.5	0.5	0.25
graphics	0.75	0.5	0.5	0.25
harrydraco	1	1	1	1
harrypotter	1	1	1	1
hiking	1	0.75	0.75	1
hp	1	0.75	1	0.5
inception	1	1	1	1
internet	0.5	0.5	0.5	0.5
ipod	1	0.5	1	0.75
jaredjensen	1	1	1	1
javascript	0.5	0.25	0.75	0.5
kirkspock	1	1	1	1
kurtblaine	1	1	1	1
make	0	0	0	0
mckaysheppard	1	1	1	1
merlin	1	1	1	1
merlinarthur	1	1	1	1
new	0	0	0	0
pairingarthureames	1	0.75	1	1
portal	1	0	0.25	0.25
power	1	0.5	0.5	0.25
programming	0.75	0.25	0.75	0.5
read	0	0	0	0
safety	1	0.75	0.5	0.25
science	1	0.5	0.75	0.5
share	0	0	0	0
sherlock	1	1	1	1
sherlockjohn	1	1	1	1
social	0.25	0.5	0.25	0.25
socialmedia	0.25	0.75	0.75	0.5

ssl	0.25	0.75	0.5	1
startrek	1	1	1	1
stevedanny	1	1	1	1
support	0.25	0.5	0.25	0.25
teachers	0.25	0.25	0.5	0.25
twitter	0.25	0.75	0.5	0.75
urban	0.5	0.25	0.25	0.5
useful	0	0	0	0
viapackratius	0	0	0	1
website	0	0.25	0.25	0

The graph in Figure 6 shows how the intuitive characterisation of tags given by humans compares with the one proposed by NTDS. The small size of most circles proves that the agreement between the humans is rather larger than before. Also, although the distribution of tags still shows the presence of umbrella terms, the proximity to the diagonal (representing total match between humans and algorithm) is much higher than before, proving that the human perception of the tags can be rather close to the NTDS cataloguing.

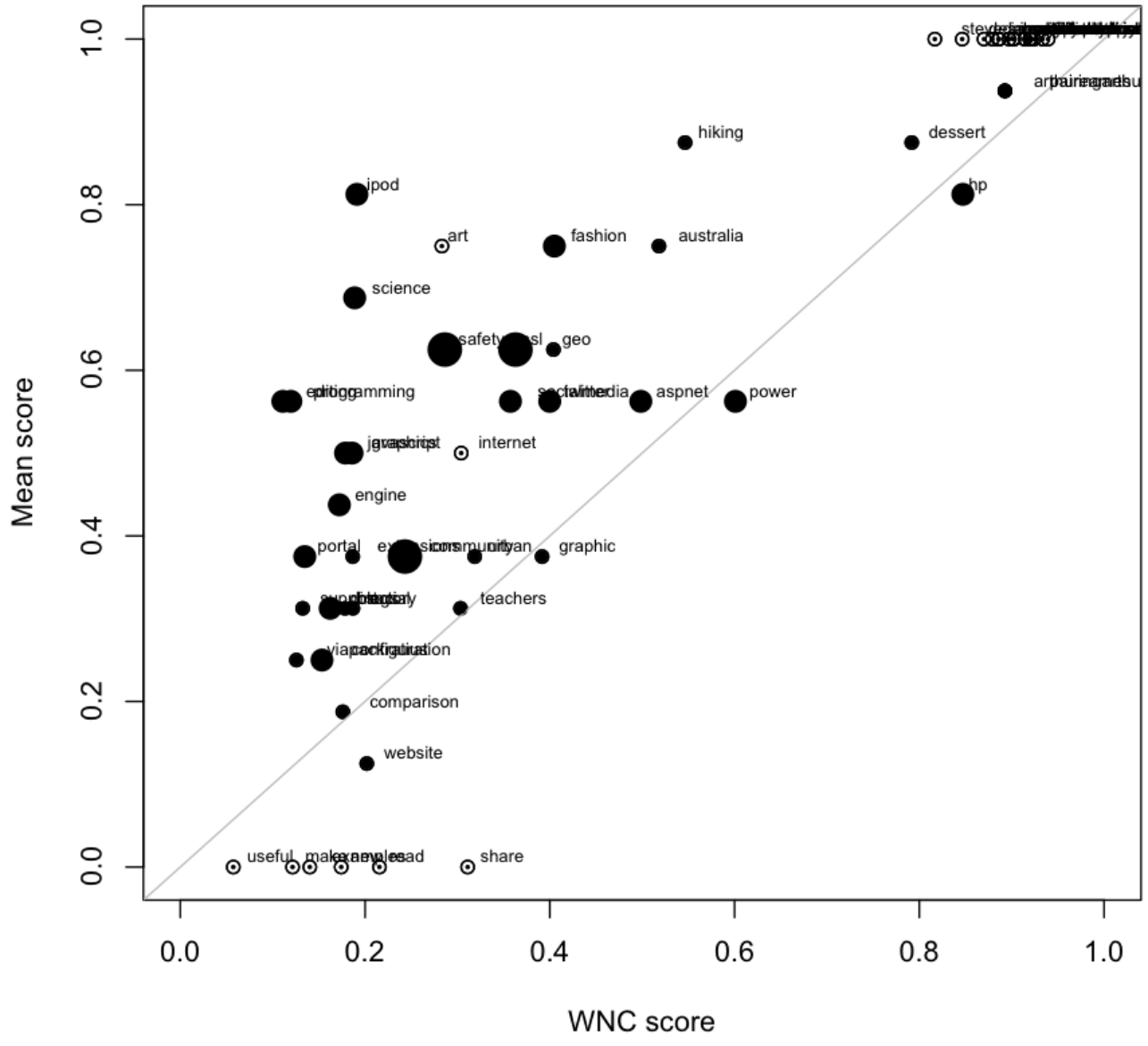

Figure 6. The tags of the third test ordered by the WNC score (X axis) and the human scores (Y axis), calculated linearizing the human's values expressed through the Likert scale between 0 and 1 and then calculating the mean. The size of the circles displays the spread in agreement between cataloguers (the larger the area, the smaller the agreement).

6. Conclusions

In this paper we introduced a classification for folksonomy tags based on two classes, *non-topical* tags and *topical* tags. We also proposed and described *Non-Topicality by Distributional Semantics* (*NTDS*), an algorithm to identify these families of tags in folksonomies based on the type of cluster which the tag ends up associated to. To test the quality of the output of NTDS, we carried out three separate tests against the outcomes of classification tasks carried out by human cataloguers.

The results show good matches for tags that are clearly topical and clearly non-topical, although the tests also showed the unforeseen emergence of a third category of tags, that we called *umbrella tags*, that have a rather widespread usage but can still be considered as topical according to many definitions and points of view.

Another unforeseen output of these tests is the unexpected but extremely frequent disagreements we found in the interpretation of the nature of the tag by human users – probably due to the intrinsic complexity of the task we assigned to them, i.e., classifying tags as either topical or non-topical – that made the comparison with the output of the NTDS algorithm rather difficult to carry out.

Of course there is still much to discover about the issue of topicality by performing additional analyses on our test set, and also by analysing data coming from other folksonomies (e.g. Flickr, Mendeley, BibSonomy). In addition, we plan to explore different mechanisms to reach agreements between humans, in order to verify how topical and non-topical tags vary in different networks and to investigate the nature of umbrella tags and their differences with topical and non-topical ones.

Notes

1. The tweet is available at https://twitter.com/BarackObama/status/426787176467034113.
2. Note that the evolution in time is a common characteristic of any social network, and it is always a crucial aspect to take into account – as social networks "change with the passing of time. People leave them while other new people appear, and with this, relationships are made and unmade. If we do not make a constant verification of our network, we can lose these valuable points of analysis as mentioned above, and end up by not seeing the changes in our network, which is never static" [35].
3. The entire dataset we gathered and other related information are available online at http://www.let.rug.nl/basile/delicious/.
4. The data collected in the tests are available at http://www.essepuntato.it/2014/ntds/tests, so that interested researchers can repeat the same experiments as desired.
5. Inter-rater agreement measures (like Cohen's K) are very useful to measure the degree of agreement among annotators but fail dramatically to capture the exact "matching" between annotations when the score distributions are severely skewed, as described by Artstein and Poesio [36]. In case of strongly connoted non-topical or topical tags most (or nearly all) annotations belongs to one single category thus measuring the inter-rater agreement in this case would result in very low Ks even if the annotations exactly match at 95%.
6. On the contrary, the tag *game* associated to a webpage containing the Flash version of PacMan would not be considered as topical, as the webpage does not talk about a game, but is itself a game.
7. In [37] are introduced two possible reasons for a misinterpretation of tags like that one we analysed: ambiguity and discrepancy on granularity. In the first case, we have an ambiguity when some terms may have multiple meanings and the choice of one of those meaning may depend on the particular background knowledge of users [38], while we observe a discrepancy on granularity when "a resource could be reasonably described by various tags, ranging from terms having a broad meaning to terms characterized by a narrow meaning" [37].
8. Queensland Museum: http://www.qm.qld.gov.au.
9. Today's Obsession: Creative Suite 5.5: http://www.printmag.com/obsessions/todays-obsession-creative-suite-5-5/.
10. We are aware that the function *getCategoriesUsedIn* should be weighted according to the actual number of occurrences of the categories appearing in the human annotations. However, since we had only four annotators, in our opinion the proposed implementation is enough to catch an intuitive level of agreement between humans.

Acknowledgements

We would like to thank Claudia Wagner for having been a patient reader of a preliminary draft of this article and for the fruitful discussions we had on the duality of tags in Twitter.

References

[1] Vander Wal T. 'Folksonomy Coinage and Definition', http://vanderwal.net/folksonomy.html (2007, accessed 20 March 2015)

[2] Bates J and Rowley J. 'Social reproduction and exclusion in subject indexing: A comparison of public library OPACs and LibraryThing folksonomy', Journal of Documentation 2011; 67(3): 431–448. DOI: 10.1108/00220411111124532

[3] Guinard D and Trifa V. 'Towards the Web of Things: Web Mashups for Embedded Devices'. In: Proceedings of the 2nd Workshop on Mashups, Enterprise Mashups and Lightweight Composition on the Web (MEM 2009), http://integror.net/mem2009/papers/paper4.pdf (2009, accessed 20 March 2015)

[4] Hendler J, Shadbolt N, Hall W, Berners-Lee T and Weitzner D. 'Web Science: an interdisciplinary approach to understanding the Web', Communications of the ACM 2008; 51(7): 60–69. DOI: 10.1145/1364782.1364798

[5] De Meo P, Nocera A, Terracina G and Ursino D. 'Recommendation of similar users, resources and social networks in a Social Internetworking Scenario', Information Sciences 2011; 181(7): 1285–1305. DOI: 10.1016/j.ins.2010.12.001

[6] Bellogín A, Cantador I and Castells P. 'A comparative study of heterogeneous item recommendations in social systems', Information Sciences 2013; 221: 142–169. DOI: 10.1016/j.ins.2012.09.039
[7] Kipp MEI. '@toread and Cool: Subjective, Affective and Associative Factors in Tagging'. In: Proceedings of the 36th annual conference of the Canadian Association for Information Science (CAIS 2008), http://www.cais-acsi.ca/proceedings/2008/kipp_2008.pdf (2008, accessed 20 March 2015)
[8] Catarino ME and Baptista AA. 'Relating Folksonomies with Dublin Core'. In: Proceeding of the 2008 DCMI International Conference on Dublin Core and Metadata Applications (DC 2008), http://dcpapers.dublincore.org/pubs/article/view/915/911 (2008, accessed 20 March 2015)
[9] Golder S and Huberman B. 'Usage patterns of collaborative tagging systems', Journal of Information Science 2006; 32(2): 198–208. DOI: 10.1177/0165551506062337
[10] Xu Z, Fu Y, Mao J and Su D. 'Towards the Semantic Web: Collaborative Tag Suggestions'. In: Proceedings of the 1st Collaborative Web Tagging Workshop, http://semanticmetadata.net/hosted/taggingws-www2006-files/13.pdf (2006, accessed 20 March 2015).
[11] Strohmaier M, Korner C and Kern R. 'Understanding why users tag: A survey of tagging motivation literature and results from an empirical study', Web Semantics: Science, Services and Agents on the World Wide Web 2012; 17: 1–11. DOI: 10.1016/j.websem.2012.09.003
[12] Baptista AA, Tonkin EL, Resmini A, Van Hooland S, Pinheiro S, Mendez E, et al. 'Kinds of Tags – Progress Report for the DC-Social Tagging Community', Podcast Presentation at the 2007 International Conference on Dublin Core and Metadata Applications (DC 2007), http://hdl.handle.net/1822/6881 (2007, accessed 20 March 2015)
[13] Dublin Core Metadata Initiative. 'Dublin Core Metadata Element Set, Version 1.1, DCMI Recommendation, 14 June 2012', http://dublincore.org/documents/dces/ (2012, accessed 20 March 2015).
[14] Sen S, Lam SK, Rashid AM, Cosley D, Frankowski D, Osterhouse J, Harper FM and Riedl J. 'Tagging, communities, vocabulary, evolution'. In: Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work (CSCW 2006). 1st ed. New York, NY, USA: ACM, 2006, pp. 181–190. DOI: 10.1145/1180875.1180904
[15] Gupta M, Li R, Yin Z and Han J. 'Survey on social tagging techniques', ACM SIGKDD Explorations Newsletter 2010; 12(1): 58–72. DOI: 10.1145/1882471.1882480
[16] Hutchins WJ. 'On the problem of "aboutness" in document analysis', Journal of Informatics 1977; 1(1): 17–35.
[17] Yang L, Sun T, Zhang M and Mei Q. 'We know what @you #tag: does the dual role affect hashtag adoption?' In: Proceedings of the 21st International Conference on World Wide Web (WWW 2012). 1st ed. New York, NY, USA: ACM, 2012, pp. 261–270. DOI: 10.1145/2187836.2187872
[18] Mika P. 'Ontologies are us: A unified model of social networks and semantics', Web Semantics: Science, Services and Agents on the World Wide Web 2006; 5: 5–15. DOI: 10.1016/j.websem.2006.11.002
[19] Specia L and Motta E. 'Integrating Folksonomies with the Semantic Web'. In: Franconi E, Kifer M and May W (eds.) Proceedings of the 4th European Semantic Web Conference (ESWC 2007). 1st ed. Berlin, Germany: Springer, 2007, pp. 624–639. DOI: 10.1007/978-3-540-72667-8_44
[20] Dublin Core Metadata Initiative 'DCMI Metadata Terms. DCMI Recommendation, 14 June 2012', http://dublincore.org/documents/dcmi-terms/ (2012, accessed 20 March 2015).
[21] Yi K and Chan LM. 'Linking folksonomy to Library of Congress subject headings: an exploratory study', Journal of Documentation 2009; 65(6): 872–900. DOI: 10.1108/00220410910998906
[22] Guarino N and Welty C. 'Evaluating ontological decisions with OntoClean', Communications of the ACM 2002; 45(2): 61–65. DOI: 10.1145/503124.503150
[23] Palma R, Zablith F, Haase P and Corcho O. 'Ontology Evolution'. In: Suárez-Figueroa MC, Gómez-Pérez A, Motta E and Gangemi A (eds.) Ontology Engineering in a Networked World. 1st ed. Berlin, Germany: Springer, 2012, pp. 235–255. DOI: 10.1007/978-3-642-24794-1_11
[24] Zablith F, Sabou M, d'Aquin M and Motta E. 'Ontology Evolution with Evolva'. In: Aroyo L, et al. (eds.) Proceedings of the 6th European Semantic Web Conference (ESWC 2009). 1st ed. Berlin, Germany: Springer, 2009, pp. 908–912. DOI: 10.1007/978-3-642-02121-3_80
[25] Deerwester S, Dumais ST, Furnas GW, Landauer TK and Harshman R. 'Indexing by Latent Semantic Analysis', Journal of the American Society for Information Science 1990; 41(6): 391–407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
[26] Burgess C and Lund K. 'Modelling parsing constraints with high-dimensional context space', Language and Cognitive Processes, 1997, XII: 1–3. DOI: 10.1080/016909697386844
[27] Blei DM, Ng AY and Jordan MI. 'Latent Dirichlet allocation',. Journal of Machine Learning Research, 2003, 3 (4–5): 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.
[28] Turney PD and Pantel P. 'From frequency to meaning: vector space models of semantics',. Journal of Artificial Intelligence Research, 2010, 37: 141-188. DOI: 10.1613/jair.2934
[29] Widdows D. 'Geometry and Meaning'. 1st ed. Stanford, CA, USA: CSLI Publication, 2004. ISBN: 1575864479

[30] Heyer G, Lauter M, Quasthoff U, Wittig T and Wolff C. 'Learning relations using collocations'. In: Proceedings of the 2nd Workshop on Ontology Learning (OL 2001), http://ceur-ws.org/Vol-38/IJCAI_2001_WS_Ontologies_Heyer_etal.pdf (2001, accessed 20 March 2015)
[31] Watts D and Strogatz S. 'Collective dynamics of "small-world" networks', Nature 1998; 393: 440–442. DOI: 10.1038/30918
[32] Dorow B, Widdows D, Ling K, Eckmann JP, Sergi D and Moses E. 'Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination', http://arxiv.org/pdf/condmat/0403693.pdf (2004, accessed 20 March 2014)
[33] Miller GA. 'WordNet: A Lexical Database for English', Communications of the ACM 1995; 38(11): 39–41. DOI: 10.1145/219717.219748
[34] Halpin H, Robu V and Shepherd H. 'The Complex Dynamics of Collaborative Tagging'. In: Proceedings of 16th International World Wide Web Conference (WWW 2007). ACM, New York, NY, USA: ACM, 2007, pp. 211–220. DOI: 10.1145/1242572.1242602
[35] Monclar R, Tecla A, Oliveira J and de Souza JM. 'MEK: Using spatial–temporal information to improve social networks and knowledge dissemination', Information Sciences 2009; 179(15): 2524–2537. DOI: 10.1016/j.ins.2009.01.032
[36] Artstein R and Poesio M. 'Inter-Coder Agreement for Computational Linguistics', Computational Linguistics 2008; 34(4): 555–596. DOI: 10.1162/coli.07-034-R2
[37] De Meo P, Quattrone G and Ursino D. 'Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies', Information Systems 2009; 34(6): 511–535. DOI: 10.1016/j.is.2009.02.004
[38] Beg MMS and Ahmad N. 'Web search enhancement by mining user actions', Information Sciences 2006; 177(23): 5203–5218. DOI: 10.1016/j.ins.2006.06.011