Measuring the Topical Specificity of Online Communities

Matthew Rowe[1], Claudia Wagner[2], Markus Strohmaier[3], and Harith Alani[4]

[1] School of Computing and Communications, Lancaster University, Lancaster, UK
[2] Institute for Information and Communication Technologies, JOANNEUM RESEARCH, Graz, Austria
[3] Knowledge Management Institute and Know-Center, Graz University of Technology,Graz, Austria
[4] Knowledge Media Institute, The Open University, Milton Keynes, UK
m.rowe@lancaster.ac.uk, claudia.wagner@joanneum.at
markus.strohmaier@tugraz.at, h.alani@open.ac.uk

Abstract. For community managers and hosts it is not only important to identify the current key topics of a community but also to assess the specificity level of the community for: a) creating sub-communities, and: b) anticipating community behaviour and topical evolution. In this paper we present an approach that empirically characterises the topical specificity of online community forums by measuring the abstraction of semantic concepts discussed within such forums. We present a range of concept abstraction measures that function over concept graphs - i.e. resource type-hierarchies and SKOS category structures - and demonstrate the efficacy of our method with an empirical evaluation using a ground truth ranking of forums. Our results show that the proposed approach outperforms a random baseline and that resource type-hierarchies work well when predicting the topical specificity of any forum with various abstraction measures.

1 Introduction

In social media applications such as message boards, online social networks or photo sharing sites, communities of users evolve around certain topics. Recent work by Belak et al. [2] examined the longitudinal changes of scientific communities and found *community drift* to be a salient factor where a given community creates new descendent communities that focus on specialised topics of the parent community. An examination of attention patterns (i.e. the factors that correlated with discussion activity and attention to content) undertaken in our prior work [13] found that the specificity of an online community forum's topic was a key feature in discerning attention patterns - e.g. in a community discussing the sport *Golf* the post had to fit the forum's topic exactly, while in a forum discussing *Work and Jobs* this was not a requirement. Recommending community forums to users who are new to a topic allows them to take advantage of the collective wisdom of the community and gain expertise and knowledge, however

recommending a community which discusses specialisations of the initial topic may overwhelm the user and a general discussion community around the topic would therefore be more appropriate.

In each of these cases (community drift, attention patterns, and community recommendation) understanding the topical specificity of a community is important for: a) tracking community focus and for new community forums to be suggested to community managers that discuss specialist topics, derived from when a community forum becomes more general in its focus; b) enabling attention-patterns of communities with the same topical specificity to be examined, and therefore the theory that arose from our prior work [13] on community specialisation correlating with attention patterns to be tested, and; c) recommending communities to novice users that are more general in the topics which they focus on, thereby alleviating the potential of overwhelming the user. Given such motivations in this paper we explore the following research question: *Can we empirically characterise how specific a given community is based on what its users discuss?*

To examine this research question we present an approach that combines concept graphs, derived from linked open data, with network-theoretic measures to gauge the abstraction level of concepts discussed by users in community forums from the Irish community message board Boards.ie.[5] Our results indicate that harnessing the linked open data graph can indeed help label the specificity of a forum based on the concepts discussed therein. Our contributions in this paper are three-fold:

1. An approach to measure forum specificity using composite functions, abstraction measures, and concept graphs.
2. Abstraction measures from network-theory that function over concept graphs.
3. Experimental assessment of the performance of different combinations of: a) composite functions, b) abstraction measures, and c) concept graphs, over a community message board platform, and a novel evaluation measure that allows for top-k level-based rankings to be assessed.

We have structured the paper as follows: Section 2 describes related work in measuring and assessing properties of online communities, and existing approaches to measure specificity and abstraction of concepts. Section 3 provides preamble of concept models used to describe online community forums. Section 4 presents our method for measuring the specificity of a community forum by using a composite function to choose the most representative concept for the forum and measuring the concept's abstraction. Section 5 details our experiments in assessing the efficacy of our approach; we explain the evaluation measures used along with the experimental setup, and demonstrate how well our method performs with respect to a random baseline and experiment permutations. Section 6 relates our work to existing related work and highlights the salient findings from this paper and plans for future work, and section 7 concludes the paper.

[5] `http://www.boards.ie`

2 Related Work

In this section we describe related work in the areas of measuring community forum properties before then describing existing work measuring concept specificity and abstraction.

2.1 Measuring Community Forum Properties

Examining the topical properties of communities has been investigated in [2] in which changes in scientific community structures are examined. One salient finding from this work, after examining the longitudinal changes of the communities, is the notion of *community shift* in which a community's topic becomes more general over time, this subsequently leads to the creation of new communities where the prior community, which became more general, is their ancestor. The closest work to ours is described in [1] where Kan et al. model conversation patterns of users on Boarrds.ie and use these patterns to characterise different community forums and hierarchically cluster them, thereby attempting to reproduce the community hierarchy structure. Our work differs, however, in that we provide an empirical assessment of the accuracy of our approach, while [1] rely on an indirect, manual inspection. Additionally, we focus on the topical qualities of forum content, complimenting the work of [1] which only uses user posting behaviour.

The behaviour of online community members was examined in [9] by measuring their behaviour along five dimensions: engagement, popularity, initiation, focus dispersion, and contribution. Rowe et al. found differences between community types (i.e. ideas, communities of practice, teams) in terms of how users behaved. The behaviour measure of focus dispersion is similar to our work as it measures, at a micro-level (i.e. user), the spread of each user in their topics. However, unlike our work it does not consider how specific individual topics are, rather their distribution per user. Term distributions are also assessed in [12] where the topics of web forums and how they change over time are visualised. Trampus and Grobelnik identified topics by choosing the term with the highest Term Frequency-Inverse Document Frequency value in a given forum. In our work we use the notion of Concept Frequency-Inverse Forum Frequency to pick out the most representative concept for a given forum, similar to [12]. Mislove et al. compared the structural properties of Flickr, LiveJournal, Orkut and YouTube [6] by examining link symmetry, power law distributions of edges and nodes, and local clustering of users. Mislove et al. found high degrees of local clustering on the different platforms which contained densely populated subgroups of similar users - i.e. shared many common connections - however the authors focussed on network-structures of social networks, ignoring content and the topical characteristics of the networks.

2.2 Measuring Specificity/Abstraction

Related to our work is research in the area of social tagging systems in which researchers have been interested in understanding the different levels of tag gen-

erality (or tag abstractness) that is essential for, amongst other things, identifying hierarchical relationships between concepts. For example, Schmitz et al. [10] suggest that if resources tagged with t_0 are often also tagged with t_1 but a large number of resources tagged with t_1 are not tagged with t_0, t_1 can be considered to subsume t_0. Heymann at al. [5] represent each tag t as a vector of resources tagged with the tag and compute the cosine similarity between these vectors. This means that they compute how similar the distributions of tags are over all resources. To create a taxonomy of tags, they sort the tags according to their closeness-centrality in the similarity graph. Benz et al. [3] and present a good overview about folksonomy-based methods to measure the level of generality of given tags and evaluate these methods by comparing them with several large-scale ontologies and taxonomies as grounded measures of word generality. Strohmaier et al. [11] present a comparative study of state-of-the-art folksonomy induction algorithms that they applied and evaluated in the context of five social tagging systems.

Unlike the above mentioned work, which aims to understand different levels of tag generality, we aim to understand different levels of community generality, and therefore specialisation. In message boards, like Boards.ie, communities form around certain tags such as sports or soccer and the aim of our work is to assess the specificity level of communities rather than assessing the specificity level of the tags around which communities are formed.

3 Preamble: Concept Models of Online Community Forums

Existing work on community forum properties examined the focus dispersion of users [9] and communities [2] without considering the specificity of the topics being discussed. As we will explain shortly concept graphs can be used to measure the level of specificity of a given community forum, however such a forum must first be represented using a model that can capture the concepts referred to within forum posts. The provided dataset for our experiments, from Boards.ie, includes a set of forums F in which posts are made. Posts are provided as a set of tuples $<u, s, t, f> \in P$, where user u posted message s at time t in forum f. The message s is composed of terms that we use to build the concept models for individual communities. The focus of a community can change and alter over time, therefore we must constrain a community's model to specific time snapshots - e.g. $t' \rightarrow t''$ where $t' < t''$. To ensure the provision of content from time-delimited forum posts we derive the set $S_f^{t't''}$ using the following construct that filters through all relevant posts' contents within the allotted time window:

$$S_f^{t't''} = \{s : <u, s, t, f> \in P, t' \leq t < t''\} \tag{1}$$

Concept models contain the distribution of concepts within a given community forum over an allotted time period - i.e. $t' \rightarrow t''$. Derivation of the concepts requires the use of concept extraction methods, we use Zemanta a third-party tool that provides a concept extraction service and is provided with the forum posts as input. Given our set of post contents, $S_f^{t't''}$, we must derive concepts

that characterise the forum in the time period. We do this by processing each post content $s \in S_f^{t't''}$ using a concept extraction tool $\Psi(s)$ to return the set of concepts related to the content of s. We build the concept model for the community by recording the frequency of concept occurrences in the input posts sets, returning $A_f^{t't''}$. This set is derived using the following construct:

$$A_f^{t't''}[c_i] = |\{c_i : c_i \in \Psi(s), s \in S_f^{t't''}\}| \quad (2)$$

4 Measuring Topical Specificity

Measuring the topical specificity of a community forum requires analysing posted content and then identifying how general or specific the concepts being discussed are. In this context, we interpret a community forum's specificity in relation to its parent forum such that the topics discussed in a child forum are a subset of those discussed in its parent (e.g. *Rugby* discusses more specialised topics than *Sports*).[6] In this section we discuss a range of abstraction measures that gauge how *abstract* a community forum's topics are. As we are interested in the *specificity* of the forum, and given that specificity is the antonym of abstraction, we take the reciprocal of the following abstraction measures ($a(c)$) for individual concepts: $1/a(c)$. In order to process the community forum f we must decide on which concept, based on those found within the forum's content, to process and return the abstraction measure for. As our abstraction measures rely on the network structures of concept graphs they can be expensive to compute, therefore we use composite functions that take a forum's set of concepts, and the frequency of concept occurrences $A_f^{t't''}$, assess each concept in the given set and returns the abstraction measure of the most representative concept. We begin this section by describing how we select which concept to return as the most representative for a forum through our composite functions, before moving on to define the abstraction measures used to assess the level of abstraction of a given concept.

4.1 Composite Functions

As mentioned previously, for a given forum f over a given time step $t' \rightarrow t''$ we are given a collection of concepts derived from posts within the window. We must decide on the best way to select from these concepts a single measure of forum specificity; we therefore provide two such functions for this task.

1. *Concept Frequency:* This function uses the frequency of the concept in the forum to pick out the most commonly discussed concept. The abstraction of the chosen concept is then measured using one of our abstraction measures - which are discussed below - and its reciprocal taken to return the specificity of the forum.

[6] It isnt necessarily the case that the more specialised forum will discuss a *single* topic (e.g. rugby could have children forums Rugby Union and Rugby League for the different codes).

2. *Concept Frequency-Inverse Forum Frequency:* This functions selects the most unique concept discussed in the forum with respect to all forums. This is a modification of the existing Term Frequency-Inverse Document Frequency measure used for term indexation. The *Concept Frequency-Inverse Forum Frequency* of each concept in a given forum is measured and the concept that returns the maximum value is chosen. The abstraction of this concept is then measured and the reciprocal of this value taken as the specificity of the forum. We define the *Concept Frequency-Inverse Forum Frequency* as follows:

$$cf{-}iff(c, f, F) = \frac{|A_f^{t't''}[c]|}{max\Big(\big\{A_f^{t't''}[c'] : c' \in A_f^{t't''}\big\}\Big)} \times \log \frac{|F|}{\big|\{f \in F : c' \in A_f^{t't''}, c' = c\}\big|} \tag{3}$$

4.2 Concept Abstraction Measures

The composite functions decide on which concept to measure based on either: a) the frequency of the concept in the forum, or b) the uniqueness of the concept with respect to the other forums. To measure concept abstraction we define five measures as follows, which either leverage the network structure surrounding a concept or use the semantics of relations in the concept graph.

Network Entropy Our first measure of concept abstraction ($a(c)$) is based on work by [3] in which tag abstraction is measured through the uniformity of co-occurences. The general premise is that a more abstract tag should co-occur with many other tags, thus producing a higher entropy - as there is more uncertainty associated with the term. In the context of our work we can also apply the same notion, however we must adapt the notion of *co-occurrence* slightly to deal with concepts. To begin with we need to define certain preamble that will allow network entropy, and the below network-theoretic measures, to be calculated, using the same definition as laid out in [4]: let $G = \{V, E, L\}$ denote a concept-network, where $c \in V$ is the set of concept nodes, $e_{cc'} \in E$ is an edge, or link, connecting $c, c' \in V$ and $lb(e_{cc'}) \in L$ denotes a label of the edge - i.e. the predicate associating c with c'. We can define the weight of the relation between two concepts c and c' by the number of times they are connected to one another in the graph: $w(c, c') = |\{e_{cc'} \in E\}|$. From this weight measurement, derived from concept co-occurrence, we then derive the conditional probability of c appearing with c' as follows, using $ego(c)$ to denote the ego-network of the concept c - i.e. the triples in the immediate vicinity of c:

$$p(c'|c) = \frac{w(c, c')}{\sum\limits_{c'' \in ego(c)} w(c, c'')} \tag{4}$$

Now that we defined the conditional probability of c appearing with another concept c', we define the network-entropy of c as follows:

$$H(c) = -\sum_{c' \in ego(c)} p(c'|c) \log p(c'|c) \tag{5}$$

Network Centrality Concepts in the semantic graph G play a role in connecting other concepts together, allowing agents using a *follow-your-nose* principle to traverse the concept space and find related terms. The importance of a concept in enabling such information flow can be gauged by its *centrality* in the network: the greater the centrality of the concept, the greater its importance. As [3] defines, the notion of centrality also allows a concept's abstraction to be measured, where the more central a concept is to the network, the greater its level of abstraction. Using this notion of centrality equating to abstraction, we provide two centrality measures as follows:

Degree-Centrality The first measure uses the degree of the concept c to assess its centrality: the greater the degree of the concept, the greater its centrality in the network. The degree of c is derived by returning the ego-centric network of c and measuring its size, we maintain directions of the edges for this measure as we are concerned with the propensity of concept c to be connected from where it appears as the subject of a triple. The cardinality of the ego-centric network is then divided by the number of concepts in the concept-network with 1 subtracted (as c cannot connect to itself):

$$Cent_D(c) = \frac{|\{c' : c' \in V, e_{cc'} \in E\}|}{|V| - 1} \tag{6}$$

Eigenvector Centrality Our second centrality measure gauges the position of the concept (c) in terms of the eigen structure of the adjacency matrix of the concept graph. The theory behind using such a measure is that the centrality of a concept depends on the centrality of those concepts with which it is connected. Let A denote the adjacency matrix of the concept network where $a_{ij} \in A$, $a_{ij} = 1$ where an edge exists between concept c_i and concept c_j and 0 otherwise. Let x_i denote the centrality score for c_i, where we define x_i as:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{|A|} a_{ij} x_j \tag{7}$$

We can rewrite Eq. 7 in a vector form such that $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ denotes the vector of centrality measures for concepts $c_1, c_2, ..., c_n$ and rearrange into a solvable form: $\mathbf{Ax} = \lambda\mathbf{x}$. The λ here corresponds to the largest eigenvector of the adjacency matrix of the concept network $\mathbf{A}$, and λ_i corresponds to the eignvector centrality score for concept c_i. Therefore by solving $\mathbf{Ax} = \lambda\mathbf{x}$ we derive the centrality scores for all concepts.

Statistical Subsumption Our next measure of concept abstraction relies on the semantics of a concept graph to identify concept subsumption. According to Schmitz et al. [10] concept c subsumes (is more general than) concept c' if $p(c|c') > \epsilon$ and $p(c'|c) < \epsilon$ for some threshold ϵ. As we are using the DBPedia graph as our knowledge base for concept relations we can exploit the semantics of the edges to detect subsumption and the hierarchical nature of the relations. For this we utilise SKOS semantics and subclass-of relations within DBPedia in

order to count how many concepts a given concept c is more general than (we use DBPedia datasets as our concept graphs which is explained in the following section).

$$SUB(c) = |\{c' : c' \in V, e_{cc'} \in E, lb(e_{cc'}) \in \{<skos{:}narrower>, <rdfs{:}subClassOf>\}| \tag{8}$$

Key Player Problem The final measure of abstraction that we use is taken from Navigli & Lapatta [7] and attempts to measure the extent to which a given node in a network is a key player in the network's topology; that is, the extent to which it is important for information flow through the network. To compute this measure we measure the shortest distance - using the Bellman-Ford algorithm - from the concept to every other concept in the network and then take the sum of the reciprocal of these distances. This sum is then normalised by the number of concepts in the network excluding the one under analysis. We define this formally as:

$$KPP(c) = \frac{\sum_{c' \in V, c \neq c'} \frac{1}{d(c, c')}}{|V| - 1} \tag{9}$$

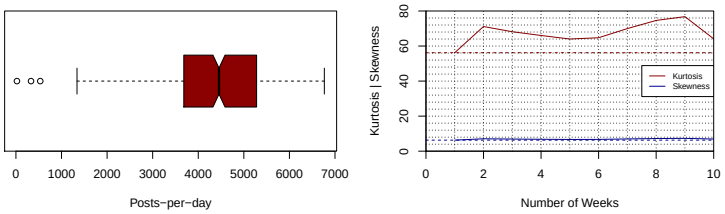
5 Experiments

In this paper we have defined how an online community forum can be modelled using the concepts discussed within its posts. We then described a method to assess the specificity of an online community forum by identifying the most representative concept and measuring the reciprocal of the concept's abstraction. Given the five different abstraction measures used and the two different composite functions, we must select the best combination to measure the specificity of a forum. Additionally, as we are using a concept graph from which to measure the abstraction of a given concept, we must also consider which source to use for the graph and examine how this affects performance.

5.1 Experimental Setup

For our experiment we needed to decide which time period to analyse - i.e. setting $t' \rightarrow t''$ - and therefore: a) where to start the period from, and b) how large the period should be. For the former point (start of the period) we counted how many posts were made every day in 2005 and found that the distribution was not normally distributed and was instead bimodal. We fitted a Gaussian mixture model using Expectation-Maximisation and found two Gaussians, thereby rendering our decision to choose a representative date based on the mean of a single Gaussian limited. We instead plotted a boxplot of the distribution, as shown in Figure 1(a), and chose the median of 4,455 posts as being the indicative point of the post distribution, we then selected the date that had 4,455 posts as our *start date*: 23/3/2005. To decide on the window size from this *start date*, we then counted how many posts were made in each forum from the *start date*

within a k-week window, and found the densities to all be normally-distributed with variance in their tails and skews. We wanted to select the most *stable* distribution of posts across the forums and therefore measured the *kurtosis* and the *skewness* of each window size's distribution - as shown in Figure 1(b). We then chose the week that produced the minimum of these measures: 1 week. By choosing this time period we are provided with reduced variation in the forum post distribution and therefore a stable picture, with no large fluctuations, of community activity.



(a) Boxplot of posts density in 2005

(b) Kurtosis and Skewness of density distributions

Fig. 1. Plots of posts-per-day distribution in 2005 (1(a)) and the distribution properties of posts-per-forum in increasing week windows from 23/3/2005 (1(b)).

The selected 1 week *experiment* period contained a total of 15,076 posts within 230 forums. We ran the text analysis tool Zemanta[7] based on prior work by Rizzo and Troncy [8], noting that this named entity recognition tool worked best on news story corpora,[8] over the post contents in the time period - 23/3/2005 $\rightarrow$ 30/3/2005 - and used the DBPedia mappings between entities and concepts to generate the concept sets: $A_f^{t't''}$. We extracted 24,752 unique entities from 15,076 posts.

Concept Graphs Mappings are required between entities and concepts as Zemanta returns DBPedia URIs which may refer to both named entities and concepts. Therefore for the former we must then identify the concepts that the entities relate to. To do this we loaded the Ontology Infobox Types and Articles Categories DBPedia datasets into Jena TDB and queried the store for mappings between an entity and: a) the class that the entity is a type of; and b) the wikipedia category that the entity is related to. We then used two graphs to assess the specificity of a forum: a) the *DBPedia Ontology Graph*, which we refer to as the *Type graph*, containing the class structure in which classes form a strict hierarchy based on `rdfs:subClassOf` relations, and; b) the *SKOS Category Graph*, which we refer to as the *Category graph*, containing the category structure from wikipedia in which categories form a loose hierarchy based on

[7] `http://www.zemanta.com/`

[8] We also note that our domain differs from that of news, but the natural language structure is similar and does not contain abbreviated forms as with Microposts.

`show:broader` relations. Our evaluation therefore, not only looks for the optimum combination of abstraction measure and composite function, but also which concept graph to use: the Type graph or the Category graph.

Table 1. Example rankings of forums in two predicted ranks from model 1 (M1) and model 2 (M2) together with the ground truth. The label function $l(.)$ returns the level of the forum from the ground truth. Our evaluation measures (Kendall τ_b and *Impurity@k*) are provided with the ordered levels as input.

	GT		**M1**		**M2**	
Rank Index	d	$l(d)$	$\hat{d_1}$	$l(\hat{d_1})$	$\hat{d_2}$	$l(\hat{d_2})$
1	a	1	c	2	a	1
2	b	1	d	2	b	1
3	c	2	g	3	c	2
4	d	2	h	3	d	2
5	e	2	a	1	f	2
6	f	2	e	2	g	3
7	g	3	i	3	e	2
8	h	3	b	1	h	3
9	i	3	j	3	i	3
10	j	3	f	2	j	3

Evaluation Measures To evaluate our approach we use the different combinations of: a) composite functions, b) abstraction measures, and c) concept graphs, to produce a predicted rank ($\hat{\boldsymbol{d}}$) - ordering the most specific forum to the most general - which is then compared against a ground truth rank ($\boldsymbol{d}$). The ground truth rank of the forums is derived from the hierarchical structure of Boards.ie which allows a given forum to be declared as either a parent or a child of another forum, thereby creating a nested structure. In this setting there are three levels that a given forum can be placed in: 1 is most specific, 3 is most general and 2 is in-between. In order to aid comprehension of our evaluation setting we present example rankings produced by two hypothetical models (M1 and M2) in Table 1 along with the ground truth (GT). We refer to this evaluation setting as *level-based ranking* as each model (M1, M2) returns a level ordering (using a label function $l(.)$) derived from the ordering of forums by their specificity values.

We use two evaluation measures for our experiments. The first measure is the Kendall τ_b coefficient which measures the difference in the number of concordant and discordant pairs and normalises this by the number of compared items - accounting for ties: -1 is a perfect negative correlation, 0 is no correlation and 1 is a perfect positive correlation. This measure yields 0.125 and 0.75 for model M1 and model M2 respectively from Table 1, indicating that M2 is better.

The second measure is a novel metric for level-based rankings called *Impurity@k* which assesses the rank up to a given point - i.e. top-k - by gauging the distance from each wrongly positioned forum to its true position in the ground truth, it is therefore equivalent to an *error measure*. The measure has a co-domain of $[0, 1]$ where 0 indicates that there are no wrongly positioned items and 1 indicates that bottom-ranked forums are ranked at the top. *Impurity@k* is derived by taking the set of outlier items (O) - derived as the set of specialised forums that are appear lower-down the rank than more general forums - and working out the distance in the rank between each outlier in the predicted rank and its true position. For model M1 from Table 1 the set of outliers contains

$O = \{a, b, f\}$ while for M2 the set contains $O = \{e\}$. For the true position we use the lowest position of a forum with the same hierarchy level as the outlier - e.g. forum a from M1 is in level 1 which has a lowest position of rank index 2 (forum b in the ground truth). We then gauge the displacement of the forum as a normalised value by setting $|F|$ as the denominator - e.g. for forum a this would be the difference between its rank index in M1 (5) and the lowest rank of a level 1 forum (2) thereby yielding 3/10 given that there are 10 forums under analysis. The normalised displacement values of each outlier are then summed and the average taken. We define this formally as:

$$impurity(k) = \frac{1}{|O|} \sum_{f \in O} \frac{|\hat{\boldsymbol{d}}^k(f) - levelrank(f, \boldsymbol{d^k}|}{|F|} \tag{10}$$

$$levelrank(f, \boldsymbol{d^k}) = \max(\{i : i = \boldsymbol{d}^k(g), l(g) = l(f), g \in F\}) \tag{11}$$

For *Impurity@k* we used six settings for k ($k \in \{1, 5, 10, 20, 50, 100\}$) and averaged the results of these values as a single measure. In doing so we concentrated on the upper-portion of the rank and therefore tested the performance of identifying topically-specific forums. For the rankings in Table 1, M1 and M2 produce *Impurity@10* values of 0.433 and 0.1 respectively (M2 is better).

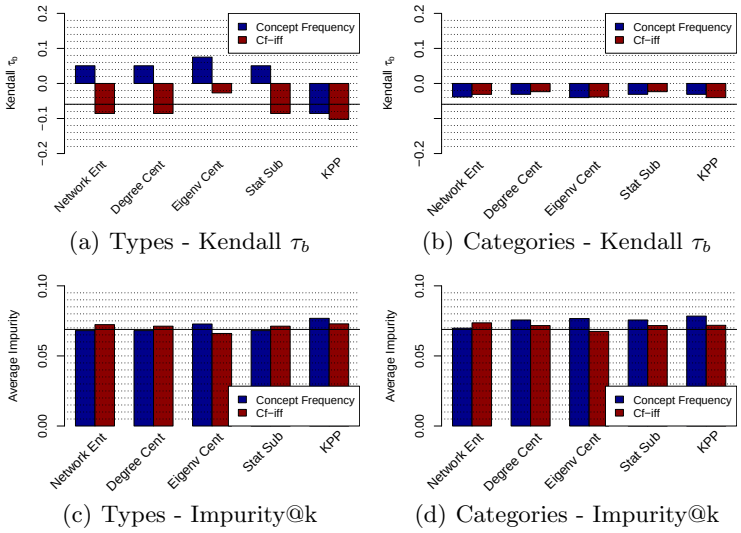
Baseline Model: Knuth Shuffle In order to aid comprehension of our results obtained using different model combinations, we compare the performance of each combination to a baseline model constructed using the Knuth Shuffle. To perform the shuffle we took the set of 230 ranked forums and iterated over the set, for each iterated forum we replaced it with a random indexed forum. Baseline measures were found to be 0.069 for *Impurity@k* and -0.0593 for Kendall τ_b.

5.2 Results

Figure 2 presents the results from different combinations of: a) composite functions, b) abstraction measures, and c) concept graphs. We see a marked difference between the performance of the Type graph (Figure 2(a)) and the Category graph (Figure 2(b)) in terms of the Kendall τ_b. We achieve the best performance when predicting the total rank using the Type graph and the Concept Frequency composite function, while using the Concept Frequency-Inverse Forum Frequency (CF-IFF) function achieves the worst performance (worse than our Knuth Shuffle baseline). This indicates that the Type graph contains sufficient information to gauge the specificity of all forums based on the classes of entities found within the forums' content. Using the frequency of the entity-types provides the best combination: achieving the best performance when using Eigenvector Centrality as the abstraction measure - we found this measure to be significantly better with the Concept Frequency function than the closest best performing combination of CF-IFF with Eigenvector Centrality when using the Type graph ($p < 0.05$ using the Sign test).

The *Impurity@k* results for the Type graph (Figure 2(c)) and the Category graph (Figure 2(d)) also show clear differences: the best performing model is the

Type graph with CF-IFF and Eigenvector Centrality (lower error than the baseline) despite this model performing poorly when predicting the total rank. The worst performing model was the Category graph, Concept Frequency and the Key Player Problem (KPP) abstraction measure, which also performed poorly when predicting the total rank (kendall τ_b). For our earlier best performing model (Type graph with Concept Frequency and Eigenvector Centrality) we do slightly worse than the random baseline, thereby failing to achieve the best performance when focussing on top-k ranks.



(a) Types - Kendall τ_b (b) Categories - Kendall τ_b

(c) Types - Impurity@k (d) Categories - Impurity@k

Fig. 2. Plots of the results obtained when measuring forum specificity using: a) the DBPedia type graph, and b) the DBPedia SKOS Category graph. The black horizontal line indicates the performance of the Knuth Shuffle random baseline.

Our results indicate that when predicting the complete ranking of communities by their topical specificity using the DBPedia Type graph and Concept Frequency yields the best model (using Eigenvector Centrality). When concentrating on forums that are focussed on a specific topic and identifying forums that are more specific than one another, then the Concept Frequency-Inverse Forum Frequency (CF-IFF) function with the Type Graph and Eigenvector centrality is best. CF-IFF returns the most unique concept for a forum with respect to other forums and using this with the Eigenvector centrality measure returns a low centrality score for any concept on the periphery of the concept graph (returning forum-specific concepts that are unique). We validated our findings using the Mann-Whitney-Wilcoxon test setting the null hypothesis that there is no difference between the specificity values attributed to forums from different levels. We achieved low p-values for the Type Graph with Concept Frequency and Network Entropy, Degree Centrality and Statistical Subsumption ($p = 0.17$, failing to reject the null hypothesis at $\alpha = 0.1$), while for Eigenvector Centrality

with CF-IFF and the Type Graph we found a significant difference between the forum level specificity values ($p < 0.1$).

Table 2 presents top-10 ranks for four model combinations (using the Type Graph as this performed best overall) indicating that different rankings are produced by the models. Similarities are evident when the same composite function is used: *Discworld* appears at the top of both abstraction measures when using Concept Frequency - indicating that the concept selected from this forum has the same specificity levels for both abstraction measures - while *Subscribers*, despite being a mid-level forum, appears towards the top rank of each abstraction measure when using CF-IFF - indicating the existence of a concept unique to this forum which shares a similar specificity level across the measures. Such qualitative analysis indicates that despite the composite functions selecting the same concept to measure the abstraction of, the measures produce, in general, different rankings based on the concept's network position.

Table 2. Forum rankings using the Type Graph and different combinations of composite functions and abstraction measures. The integers in parentheses represent the level of the forum on Boards.ie: 1=most specific, 3= most general.

Concept Frequency		**CF-IFF**	
Network Entropy	**Eigenv' Cent'**	**Network Entropy**	**Eigenv' Cent'**
Discworld (1)	Discworld (1)	Languages (1)	Magic the Gathering (1)
The Cuckoo's Nest (2)	Angling (2)	Hunting (1)	Subscribers (2)
Models (2)	Paganism (1)	File Exchange (2)	Unreal (2)
Slydice Specials (1)	Feedback (2)	Game Threads (1)	LAN Parties (2)
Battlestar Galactica (1)	Personal Issues (2)	Magic the Gathering (1)	World of Warcraft (1)
FS Motors (1)	Mythology (2)	Bangbus (1)	Role Playing (2)
Gadgets (1)	Films (1)	Biology & Medicine (2)	Midwest (2)
FS Music Equipment (1)	Business Managem' (1)	Snooker & Pool (2)	Game Threads (1)
Pro Evolution Soccer (2)	Xbox (1)	Subscribers (2)	GAA (2)
Call of Duty (2)	Help Desk (2)	HE Video Players (1)	Midlands (2)
Anime & Manga (2)	DIT (2)	Discworld (1)	Discworld (1)

6 Discussions and Future Work

Existing research on social tagging systems [10, 5] attempts to assess the specificity of a tag in order to build tag hierarchies. Our work is analogous to tag hierarchy construction as it will enable hierarchies of communities to be constructed in a similar vein to [1]. Our future work will compare results for hierarchical clustering of the forums using specificity values from the best performing model - i.e. Eigenvector Centrality with the Concept Frequency composite function and the Type graph - with the clustering from [1] in order to test how well our measures replicate forum hierarchies and structures. When exploring the longitudinal behaviour of scientific communities Belak et al. [2] identified *community shift* as being a prevalent phenomena where a community spawns new communities that are specialisations of their ancestor (parent community). Our work contributes to such explorations by performing specificity analysis of online community forums: if one can track the specificity of a community over time, then one can identify topic shift and inform community managers as to

which new topics could be used for community forums, identifying such events based on the increased generality of a community's topic.

In our prior work [13] we theorised that communities which focused on specific topics showed similar attention patterns - where a post starting a discussion thread had to match the community's topic of interest - while these specific topic communities differed from general discussion communities. The work presented in this paper provides the necessary means for empirically measuring the topical specificity of communities on Boards.ie and other community measure boards. Therefore our future work will involve grouping communities by their topical specificity - measured using Eigenvector Centrality as our abstraction measure, Concept Frequency as our composite function, and the DBPedia Type graph as our concept graph - and examining the attention patterns of specific communities vs general communities, thereby proving, or disproving, our earlier theory from [13]. In this paper we have considered a *semantic approach* to measure the topical specificity of online community forums, however there is the potential to also examine an alternative purely *social approach*: for instance, based on the notion of *Statistical Subsumption* which we explored as one of our abstraction measures, one could identify forum f_a as being more general than forum f_b if the set of authors who created posts on f_b is a subset of the authors who authored posts in f_a. Such insights and potentials for future work have been afforded as a result of the work discussed within this paper.

7 Conclusions

In this paper we presented an approach to measure the topical specificity of online community forums that used abstraction measures which functioned over concept graphs and composite functions to return a representative concept for a community, and thereby its specificity level. Motivated by our research question (*Can we empirically characterise how specific a given community is based on what its users discuss?*) the empirical assessment of forum specificity through our experiments showed the divergent performance between different composite functions, abstraction measures and concept graphs, where the use of a resource type-graph derived from the DBPedia Type Ontology provided a useful resource for predicting a complete ranking of forums by their specificity levels, outperforming the SKOS Category structure. We also found that using the Eigenvector Centrality measure and the Concept Frequency-Inverse Forum Frequency function provided the best combination for identifying differences in topic-specific communities to be discerned - this latter assessment being measured using our novel evaluation metric $Impurity@k$ that accounts for top-k ranked levels.

The results and findings from this work will inform our future work on examining attention patterns of online communities, and also enable the longitudinal assessment and tracking of a community forum's topical specificity, thereby allowing new communities to be recommended to managers based on topical generalisation - a natural life-cycle of communities as put forward by [1].

Acknowledgment

This work was supported in part by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner.

References

1. Andrey Kan, Jeffrey Chan, Conor Hayes, Bernie Hogan, James Bailey, Christpher Leckie. A Time Decoupling Approach for Studying Forum Dynamics. *World Wide Web Internet And Web Information Systems*, In press:1–24, 2011.
2. VÌaclav Belak, Marcel Karnstedt, and Conor Hayes. Life-cycles and mutual effects of scientific communities. *Procedia - Social and Behavioral Sciences*, 22(0):37 – 48, 2011.
3. Dominik Benz, Christian Körner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. One tag to bind them all: measuring term abstractness in social metadata. In *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II*, ESWC'11, pages 360–374, Berlin, Heidelberg, 2011. Springer-Verlag.
4. Christophe Guéret, Paul T. Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In *ESWC*, pages 87–102, 2012.
5. Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department, April 2006.
6. Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *SIGCOMM conference on Internet measurement*, IMC '07, pages 29–42, 2007.
7. Roberto Navigli and Mirella Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(4):678–692, 2010.
8. G. Rizzo and R. Troncy. Nerd: evaluating named entity recognition tools in the web of data. In *Workshop on Web Scale Knowledge Extraction (WEKEX'11), Bonn, Germany*, pages 1–16, 2011.
9. M. Rowe, M. Fernandez, H. Alani, I. Ronen, C. Hayes, and M. Karnstedt. Behaviour analysis across different types of enterprise online communities. In *ACM Web Science Conference*, 2012.
10. Patrick Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland, May 2006.
11. Markus Strohmaier, Denis Helic, Dominik Benz, Christian Körner, and Roman Kern. Evaluation of folksonomy induction algorithms. *Transactions on Intelligent Systems and Technology*, 2012.
12. M. Trampuš and M. Grobelnik. Visualization of online discussion forums. In *Workshop on Pattern Analysis Applications*, 2010.
13. C. Wagner, M. Rowe, M. Strohmaier, and H. Alani. Ignorance isn't bliss: an empirical analysis of attention patterns in online communities. In *AES Conference on Social Computing*, 2012.