Automatic Classification of Documents in Cold-start Scenarios

Ricardo Kawase
Leibniz University of Hanover
& L3S Research Center
Appelstrasse 9, 30167
Hannover, Germany
kawase@L3S.de

Marco Fisichella
Leibniz University of Hanover
& L3S Research Center
Appelstrasse 9, 30167
Hannover, Germany
fisichella@L3S.de

Bernardo Pereira Nunes
Leibniz University of Hanover
& L3S Research Center
Appelstrasse 9, 30167
Hannover, Germany
nunes@L3S.de

Kyung-Hun Ha
ESCP Europe
Wirtschaftshochschule Berlin
Heubnerweg 8-10, 14059
Berlin, Germany
Kyung-Hun.Ha@escpeurope.eu

Markus Bick
ESCP Europe
Wirtschaftshochschule Berlin
Heubnerweg 8-10, 14059
Berlin, Germany
Markus.Bick@escpeurope.eu

ABSTRACT

Document classification is key to ensuring quality of any digital library. However, classifying documents is a very time-consuming task. In addition, few or none of the documents in a newly created repository are classified. The non-classification of documents not only prevents users from finding information but also hinders the system's aptitude to recommend relevant items. Moreover, the lack of classified documents prevents any kind of machine learning algorithm to automatically annotate these items. In this work, we propose a novel approach to automatically classifying documents that differs from previous works in the sense that it exploits the wisdom of the crowds available on the Web. Our proposed strategy adapts an automatic tagging approach combined with a straightforward matching algorithm to classify documents in a given domain classification. To validate our findings, we compared our methods against the existing and performed a user evaluation with 61 participants to estimate the quality of the classifications. Results show that, in 72% of the cases, the automatic classification is relevant and well accepted by participants. In conclusion, automatic classification can facilitate access to relevant documents.

Categories and Subject Descriptors

H.3.2 [**Information Search and Retrieval**]: Information Storage—*Record classification*; H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Languages

Keywords

Information Retrieval, Automatic Classification, Cold-start, User Evaluation, Digital Libraries

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WIMS'13, June 12-14, 2013 Madrid, Spain
Copyright ©2013 ACM 978-1-4503-1850-1/13/06 ... $10.00

1. INTRODUCTION

Nowadays, the World Wide Web is the largest source of information. In the last years, every knowledge repository has moved their resources to online digital repositories. Consequently, the number of specific online disciplinary repositories has also increased significantly. Currently, online educational digital libraries are being deployed for all range of topics. The goal is to support learners to easily find relevant material on a particular topic. Since search engines as Google[1], Bing[2] and Yahoo[3], to name but a few, dictate information retrieval in the Web, digital libraries must offer an attractive differential for the users. The differential offered by these libraries comes in the means of focused topics, high quality resources and easy retrieval.

Since the catch up of the Open Archives Initiative[4], plenty of data is freely available. Through utilization of the OAI-PMH protocol, a digital library can list the contents of several external repositories. However, digital libraries that rely on external content usually suffer with the issue of assuring content quality since they do not own the actual documents. Nevertheless, these gatherers also need to maintain a minimal threshold of quality, accessibility, and usability. Thus, it is crucial for any digital library to evaluate each new resource they receive by judging its quality and relevance to the collection. In most cases, evaluations are manually performed by curators who are familiar with the scope of the collection. However, as the amount of available content rises exponentially, it becomes an unfeasible task for humans. This issue

[1] http://www.google.com
[2] http://www.bing.com
[3] http://www.yahoo.com
[4] http://www.openarchives.org

is even more problematic for the cases of Open Archives, where a new repository may be added to the library with thousands of new documents at once.

To overcome this information overload problem and to maintain the quality of the collections, there is a lot of research focusing on the quality assurance of resources as well as facilitation of access to information. For example, state-of-the-art work from Bethard et al. [1] proposed methods to automatically identify out-of-scope resources. In another direction, several other works approach the problem of automatically classifying documents [10, 18, 15, 20], thus identifying if one document belongs or not to a collection.

For the vast majority of previous works in this area, the methods are basically built on top of machine learning strategies. They propose different solutions for the classic text classification problem that have always the basic assumption of an existing training dataset. The work we present in this paper completely differs from previous works in the area, as we take into consideration the deeper problem of there being no prior information on the corpus of the collection. In a few words, our proposed task is to classify an entirely unclassified collection. This issue is not exclusive of digital libraries. By modeling this problem as a recommendation task, the goal is to recommend a category to a document that has no prior connection with the collection (the so called cold-start problem).

In this work, we propose to automatically classify learning objects by exploiting the content from different but similar resources found outside the boundaries of a single content repository. Our automatic classifying method is an extension of the state-of-the-art α-*TaggingLDA* for automatic tagging [4], which is based on the probabilistic topic model Latent Dirichlet Allocation [2].

The main difference between tagging and classifying is that, the task of tagging documents is not limited to a restricted vocabulary. Thus, there is no completely right or completely wrong answer. Tags may be not completely relevant to a document but yet they always attach additional information to the resource. On the other hand, the classification task requires a more precise and focused analysis since the outcomes must be within the boundaries of a fixed vocabulary. Differently from tags, categorization has a binary assessment that is either right or wrong. Additionally, misclassification has a greater impact for the user than misplaced tags. Under the user's perspective, a misclassified document may bias the reader towards misunderstanding of the content and even, totally preclude the document from being discovered.

Since we are dealing with a new, unclassified digital library, we evaluated our outcomes with two user studies. Additionally, we validated our method with an existing benchmark for manually categorized repositories. In doing so, we address the following research questions:

Q1: From the user's perspective, how relevant for the document is automatic classification?

Q2: To what extent the classifications assigned automatically agree with those given by human judgments?

Q3: Are automatic classifications useful and effective in other domains?

Furthermore, the thorough analysis of the participants' behaviors during the classification tasks serves as an evaluation of the quality of the documents in the repository and the quality of the proposed domain classification.

The contributions of this work are:

- An automatic classifying approach to efficiently addressing the cold-start problem in digital libraries relying on content from resources in an auxiliary domain, i.e., one that lies outside of the content repository of the unclassified document.
- An evaluation of our approach through a user study involving 61 participants in an online setting, with real-world data, and experiments utilizing a general-domain benchmark dataset.

The rest of this paper is organized as follows. In Section 2, we will first present related works in the area of automatic text classification. In Section 3, we describe the basis of our approach and how it has already been validated. In Section 4, we describe in details the strategy used to create the automatic classifier. Section 5 describes the experiences on putting together a new digital library and a new domain classification. In Section 6, we present a two-part evaluation to measure the effectiveness of our method and in Section 7, we describe the evaluation results. In Section 8, we validate our method based on a benchmark dataset. Finally, in Section 9, we conclude and discuss our findings and the directions for future work.

2. RELATED WORK

Much research has been done to improve the task of automatically classifying documents. Typically, the classification task can be understood in two ways. First, in the sense of assigning classes (predefined terms) to a document. Second, as we approached in this work, strictly grouping documents into one class. In this area, important research has been conducted by Fisichella et al. in [7]; the authors assign each document to one class, using a soft clustering algorithm, which is described by a set of terms. In both cases, the final goal is to improve organization and information retrieval. A great part of the literature on text classification is based on machine learning approaches and rely on dimensionality reduction [17] or on probabilistic topic models [4]. These strategies begin with a large set of manually annotated documents (positive examples of classification) where algorithms find existing patterns in documents in each class. Then, in a second step, these patterns are automatically identified in non-classified documents [10, 18, 15].

Although there is vast literature in the area, the basic idea is immutable. Each algorithm exploits different features and implements unique strategies to identify patterns that can later be used to classify new documents.

In many studies, the well accepted approach to begin with text classification is TF-IDF weighting [18, 11, 13, 19]. This well known strategy turns documents into a list of weighted terms that facilitates the representation of the documents. It

relies on the assumption that the most representative terms of a document occur many times in the document's text and, at the same time, occur only in a small set of the available documents. To the best of our knowledge, the most successful approaches for automatic classification are based on TF-IDF, usually combined with support vector machine (SVM) classifiers [11, 19, 1]. Standard SVM approaches try to predict, from input data, two possible classes maximizing their margin.

In all visited previous works, there is always the assumption of an existing training data. Our work distinguishes from the previous work on document classification in two ways. First, we do not build upon any pre-existing human annotated data. Second, we do not base our strategy on incremental machine learning algorithms. Since there is no training data that feed the method with confident positive examples, there is no learning strategy to build upon. As exposed in the following sections, our proposed method is composed of strategies that exploit existing knowledge of outside repositories, combined with the wisdom of the crowds and a straightforward heuristic approach. Our approach relies on probabilistic topic models, in particular on Latent Dirichlet Allocation (LDA), as described and referenced in section 3. Nevertheless, we performed an online user-study to collect enough human annotated data in order to evaluate our automatic classifier.

3. BACKGROUND

Before explaining our approach for enhancing learning objects metadata with tags, we first present in this section some terminology and background about the concepts discussed in this paper.

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2] is a generative probabilistic model for collections of discrete data such as text corpora. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over terms. More formally, assume that a text collection consists of a set of documents D. Furthermore, consider the set of topics Z, the distribution $P(z \mid d)$ over topics $z \in Z$ in a particular document $d \in D$ and the probability distribution $P(t \mid z)$ over terms $t \in T$ given topic $z \in Z$, where T is the set of terms. Each term $t_i \in T$ in a document (where the index refers to the ith term token) is generated by first sampling a topic from the topic distribution, then choosing a term from the topic-term distribution. We write $P(z_i = j)$ as the probability that the jth topic was sampled for the ith term token and $P(t_i \mid z_i = j)$ as the probability of term t_i under topic j. The model specifies the following distribution over terms within a document:

$$P(t_i \mid d) = \sum_z P(t_i \mid z_i = j) P(z_i = j \mid d) \qquad (1)$$

where $|Z|$ is the number of topics. $P(t \mid z = j)$ and $P(z \mid d)$ indicate which terms are important for which topic and which topics are important for a particular document, respectively.

In LDA the goal is to estimate the topic-term distribution $P(t \mid z)$ and the document-topic distribution $P(z \mid d)$, these distributions are sampled from Dirichlet distributions.

There are several methods developed for making inference in LDA such as variational expectation maximization [2], expectation propagation [3], and Gibbs sampling [8].

The Gibbs Sampling algorithm, for example, considers each term token in the text collection in turn, and estimates the probability of assigning the current term token to each topic, conditioned on the topic assignments to all other term tokens. From this conditional distribution, a topic is sampled and stored as the new topic assignment for this term token. This conditional distribution can be written as $P(z_i = j \mid t_i, d_i, z_{-i})$, and calculated by [8]:

$$P(z_i = j \mid t_i, d_i, z_{-i}) \propto \frac{C^{TZ}_{t_i j} + \beta}{\sum_t C^{TZ}_{tj} + |T|\beta} \frac{C^{DZ}_{d_i j} + \alpha}{\sum_z + |Z|\alpha} \qquad (2)$$

where C^{ZT} and C^{DZ} are matrices of counts with dimensions $|T| \times |Z|$ and $|D| \times |Z|$ respectively; C^{ZT}_{tj} contains the number of times term t is assigned to topic j, not including the current instance i, and C^{DZ}_{dj} contains the number of times topic j is assigned to some term token in document d, not including the current instance i. $z_i = j$ represents the topic assignment of token i to topic j, z_{-i} represents all topic-term and document-topic assignments except the current assignment of z_i to term t_i, and α and β are the (symmetric) hyper-parameters for the Dirichlet priors. Based on the count matrices the posterior probabilities in Equation 1 can be estimated as follows:

$$P(t_i \mid z_i = j) = \frac{C^{TZ}_{t_i j} + \beta}{\sum_t C^{TZ}_{tj} + |T|\beta} \qquad (3)$$

$$P(z_i = j \mid d) = \frac{C^{DZ}_{d_i j} + \alpha}{\sum_z + |Z|\alpha} \qquad (4)$$

LDA is an intensively studied model and its performance compares favorably to other known information text retrieval techniques. In addition to the large number of applications in this field, LDA has also been applied to several other problem scenarios, including entity resolution [6], image processing [12, 14], fraud detection [21], and many more.

3.2 α-TaggingLDA

α-TaggingLDA is a state-of-the-art LDA-based approach for automatic tagging introduced by Diaz-Aviles et al. [5]. *α-TaggingLDA* is designed to overcome new item cold-start problems by exploiting content of resources, without relying on collaborative interactions.

An overview of the *α-TaggingLDA* method is shown in the upper part of Figure 1. In order to illustrate the method with an example, consider a novel LO entitled *Knowledge Technologies in Context.* This resource is new to the collaborative learning system and does not have any tag annotations assigned. The absence of tags makes it difficult for the

system to consider it as candidate for recommendations, for instance.

α-*TaggingLDA* first extracts relevant *textual content* from the LO, such as the title, description or metadata (e.g., author), and creates a document denoted as d_{LO}. Then, the LO is associated with a set of 'similar' documents, which we refer to as an *ad hoc corpus* for the LO, represented as $corpus_{LO}$.

Note that the α-*TaggingLDA* method does not impose any restriction on the similarity measure used to associate the corpus with the LO. The similarity measure could be specified based on the nature of the resources, (e.g., text documents, multimedia items) and the textual content or metadata available. For example, a particular implementation might rely upon a computationally inexpensive similarity measure or on a more complex clustering algorithm.

In our particular example, the title of the LO is used to query an Internet search engine in order to retrieve the title and snippets of the n relevant results ($n = 4$, in this case). This subset corresponds to $corpus_{LO}$.

The LO's textual content is extracted and the subset of the top n results constitutes the text collection $D = \{d_{LO}\} \cup corpus_{LO}$, which is input for LDA, together with the number of topics required. In this example, the number of topics is set to two, i.e., $|Z| = 2$. The set of tags to be used to annotate the LO is denoted as $TopN_{tags}(LO)$, and its size is set to six for this particular case, i.e., $|TopN_{tags}(LO)| = 6$.

Table 1 presents an example of the output produced by LDA according to the setting described above. Topics are ordered based on the document-topic distribution $P(z \mid d)$, and within each topic, terms are ranked based on the topic-term $P(t \mid z)$ distribution.

For the construction of the final set of tags $TopN_{tags}(LO)$, α-*TaggingLDA* selects the first candidate tag from $Topic_1$'s top terms, the second tag from $Topic_2$'s top terms, the third tag, again from $Topic_1$'s top terms, and so forth. The final list of tag annotations for the LO in our example corresponds to $TopN_{tags}(LO) = \{$ *technologies, phenomena, software, work, ecosystems, business* $\}$. For the details of this strategy, we refer the reader to the work done by Diaz, et.al. [5].

Table 1: Example of two topics output by LDA. Topics are ordered based on the document-topic distribution $P(z \mid d)$, and within each topic, terms are ranked based on the topic-term $P(t \mid z)$ distribution.

$\boldsymbol{Topic_1}$		$\boldsymbol{Topic_2}$	
$P(z = 1 \mid d_{LO}) = 0.70$		$P(z = 2 \mid d_{LO}) = 0.30$	
Term t	$P(t \mid z = 1)$	Term t	$P(t \mid z = 2)$
technologies	0.45	phenomena	0.33
software	0.25	work	0.28
ecosystems	0.16	business	0.19
systems	0.11	researchers	0.15
representation	0.03	vendors	0.04
interpretation	0.01	people	0.01

3.3 α-TaggingLDA Evaluation

Previously to this work, we have evaluated the α-*TaggingLDA* method for automatic tagging of learning objects[4].

We have empirically demonstrated through a series of evaluations that the α-*TaggingLDA* method produces quality metadata enhancement for learning objects. The evaluation compared the automatically generated tags against existing tag annotations performed by the authors. Furthermore, a user study compared the participants' preference for automatically produced tags against the authors' tags. Finally, the evaluation also demonstrated that α-*TaggingLDA* tags are the best candidate terms for assisting users in the tagging process.

The outcomes of evaluations with over 100 participants showed an agreement of 38.4% of the automatically generated tags with those provided by the participants. More notable was the participants' preference for the automatically generated tags (67.5%) over the experts' tags (32.5%).

In the end, the most important consideration was the potential benefits produced by information delivered through the automatic tagging method. We build upon this information to deliver the automatic classification method.

4. TAG-BASED DOMAIN CLASSIFIER

On top of the automatic tagging method presented in the previous section, we added a new layer to identify which is the most probable category a document belongs to. The classification layer uses two different inputs: first, a ranked list of keywords that describes the resource to be classified; second, a list of domains to which the document can belong, with a list of keywords describing each domain. For the first input, as previously described, α-TaggingLDA provides a ranked list of tags that, to some extent, represents the main concepts in a document.

For the second, the list of topics used by the library and a few keywords that best describe the topic are required. Describing topics with keywords is a light-weight task when compared to the manual assignment of categories for each document in a collection.

With these two inputs, the classification method assigns scores for each match found between the document's list of keywords and the domain's keywords. Since the document's keywords are already properly ranked, we apply a linear decay to the matching-score. This means that the domain's keyword that matches the first document's keywords has a greater score than those matching the document's keywords that are more highly positioned in the ranking. After the matching process, we compute the sum of the scores of each topic, assigning the top scoring to the document. The pseudocode (Algorithm 1) depicts the matching method.

We configured the α-*TaggingLDA* to return a maximum of 100 terms for each document. During the classification matching, if no correspondences were found the document was declared unclassified. To evaluate the proposed method (refer to Section 6), we utilized the OpenScout[5] project

[5]http://learn.openscout.net

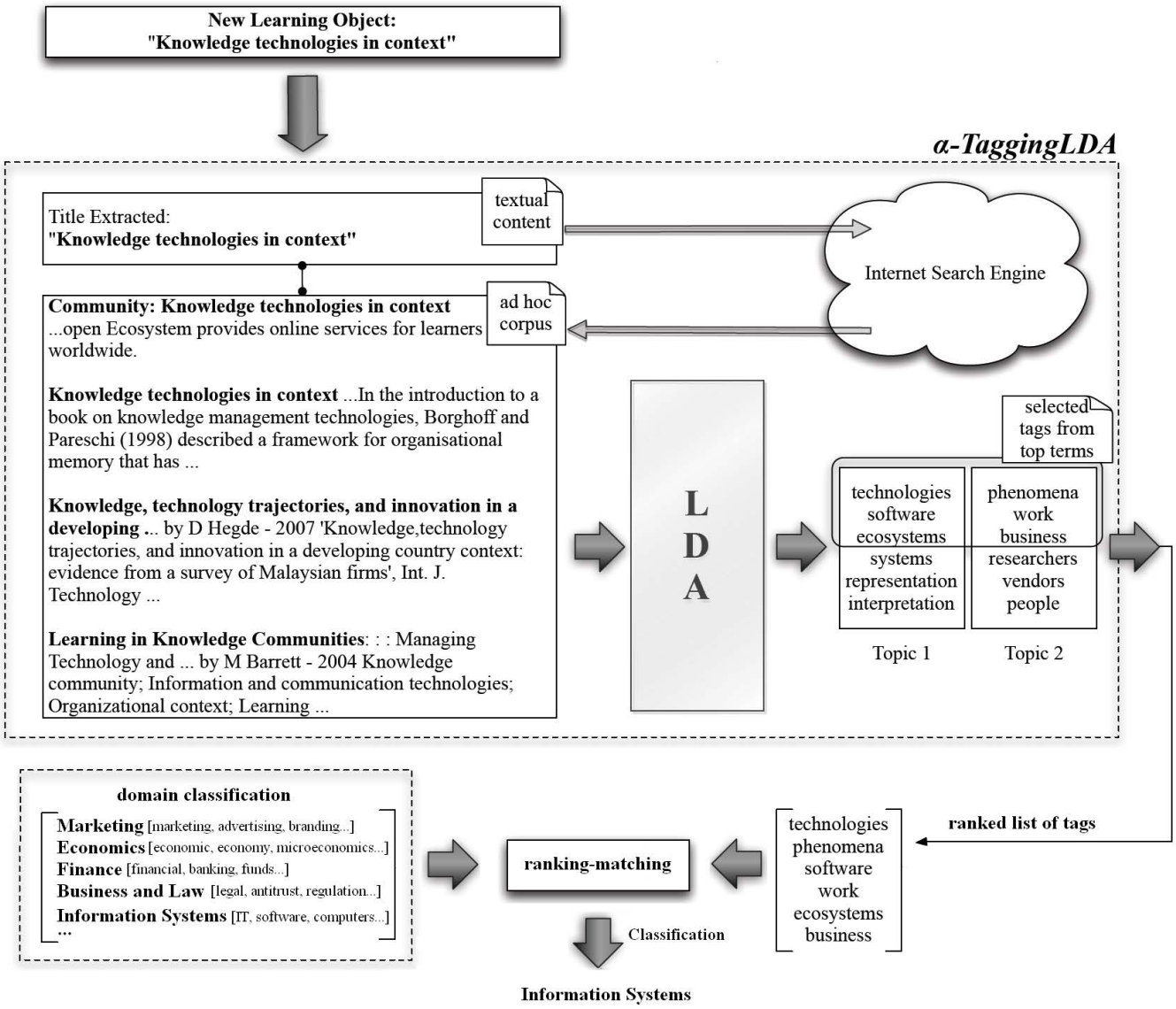

Figure 1: α-*TaggingLDA* is applied to annotate a new LO, *Knowledge Technologies in Context*, with a list of six tags: $TopN_{tags}(LO) = \{$ *technologies, phenomena, software, work, ecosystems, business* $\}$, based on two LDA topics. The ranked tag list is then matched against the domain classification. The ranking is transferred to domain classification and the top ranked category is revealed.

repository, a new digital library in the area of business and management that covers numerous topics. The project has its own domain classification that was proposed by experts in the field. Fortunately, we had access to the same experts and asked them to build a list of keywords describing each domain which we explain in Section 5.

4.1 Baseline

In order to draw a comparison of our approach with existing strategies, we chose successful well-known methods used in text classification, as presented in Section 2. First, we calculated the TF-IDF values for all words in each document within the corpus. For each document, we removed from the text words with less than 2 characters and words consisting of numbers because such terms were not useful when determining the category of an article; additionally, we removed the punctuation marks (e.g. –, ?, %, /, !, etc.) from the words and combined the remaining parts. Finally, we removed stop words and applied stemming.

For each article, we stored in one vector the top 15 remaining terms according to the highest TF-IDF values. We used such vectors of words to represent a document as surrogate; then, we assigned the classification by computing the similarity (Jaccard) of the TF-IDF results with the relevant keywords of each domain. In addition, we also performed the computation by using our proposed matching method previously presented. In Section 6, we will evaluate three distinct strategies:

Table 2: The domain classification of the OpenScout repository and the respective examples of most relevant keywords.

Domains	Relevant Keywords
Organizational Behavior and Leadership	organizational,behavior,leadership,negotiation,team,culture...
Decision Sciences	decision,risk,forecasting,operation,modeling,optimization...
Marketing	marketing,advertising,advertisement,branding,b2b,communication...
Economics	economics,economy,microeconomics,exchange,interest,rate,inflation...
Finance	finance,financial,banking,funds,capital,cash,flow,value,equity,debt...
Strategy and Corporate Social Responsibility	strategy,responsibility,society,sustainability,innovation,ethics,regulation...
Accounting and Controlling	accounting,controlling,balance,budgets,bookkeeping,budgeting...
Management Information Systems	management,information,system,IT,data,computer,computation...
Technology and Operations Management	technology,operation,ebusiness,egovernment,ecommerce,outsourcing...
Entrepreneurship	entrepreneurship,entrepreneurs,start-up,opportunity,business...
Human Resource Management	resources,management,career,competence,employee,training,relation...
Language and Communication	languages,communication,message,grammar,nonverbal,verbal...
Project Management	management,monitoring,report,planning,organizing,securing...
Business and Law	law,legal,antitrust,regulation,contract,formation,litigation...
Others	-

- TF-IDF + Jaccard
- TF-IDF + Matching
- α-TaggingLDA + Matching

5. DOMAIN CLASSIFICATION

OpenScout[6] is an EU co-funded project which aims at providing skill-and-competence-based search and retrieval Web Services that enable users to easily find, access, use, and exchange open content for management education and training. Therefore, the project not only connects leading European Open Education Resources (OER) repositories but integrates its search services into existing learning suites. Within the project, a management-related domain classification was developed (see Table 2) in order to support the learner while searching for appropriate learning resources that belong to a specific domain, e.g. marketing or finance.

Additionally, each identified domain was enriched by a list of the most important keywords describing the domain as accurately as possible. A step-by-step approach was used to develop the new OpenScout domain classification and its corresponding keywords.

In a first major step, a focus group was organized and moderated by one of the OpenScout project coordinators who has major experience in managing this form of group discussion. Focus group participants consisted of a sample of ten domain experts from Higher Education, Business Schools and Small-Medium Enterprises (SME), including two professors, six researchers, and two professionals to generate an initial domain classification based on experience and academic literature. After further in-depth discussions, and comparison with already existing domain classifications of other academic institutions, only those terms that best fit management education and the underlying project goals were finally retained by the focus group, yielding 15 fundamental domains.

A pretest with domain experts from higher learning institutions INSEAD[7], BRUNEL[8], EFMD[9] and VMU[10] was conducted to assess the content of the domain classification and to ensure content validity. Those terms that best fit management education in general, hence the content of the learning resources, were retained by the experts for the final domain classification.

As stated above, each domain was enriched by a list of main keywords. In a second major step, eight researchers from the ESCP Europe Business School[11], with different research focus and knowledge about certain domains, were asked to provide a list of eight to ten terms that best fit their domains. Participants had completed different diploma studies in Germany, the USA, UK, Australia, or China and had an average of two years of work experience at the university;

[6]http://openscout.net

Algorithm 1: Pseudocode for keyword-term matching method.

```
1  begin
2      for each document do
3          Get top N α-TaggingLDA keywords;
4          KeywordIndex=0; for each keywords do
5              KeywordIndex++; for each domain do
6                  Get domain's terms;
7                  for each domain's terms do
8                      if keyword == term then
9                          domain-score += 1/KeywordIndex;
10     return top scoring domain;
```

[7]http://www.insead.edu/home
[8]http://www.brunel.ac.uk
[9]http://www.efmd.org
[10]http://www.vdu.lt
[11]http://www.escpeurope.eu

three of them had also been previously employed full-time in several industries. Looking at the resulting keywords of each domain, all experts emphasized that they can only provide a subjective assessment as each domain represents a broad field of knowledge. However, due to their long years of experience and ongoing education in their respective field of knowledge, these experts fulfill the necessary criteria for providing the most relevant keywords.

6. EVALUATION

In this section, we measure the benefits of automatic classification for an unclassified digital library. To answer the research questions presented in Section 1, we conducted two distinct user studies. The rest of the section describes each evaluation setting.

6.1 Dataset

We based our experiments on a dataset sampled from the OpenScout project collection [16]. According to the Open Archives Initiative, the project gathers metadata information from learning resources located at different learning content repositories. For our evaluation, we selected all documents whose language was English. In total, we collected 7,750 items that should be classified under one of the 15 categories. At the time of data collection, none of the items had any information about the classification. Each document was then subjected to automatic classification by each one of the methods, namely, TF-IDF+Jaccard, TF-IDF+Matching, and α-*TaggingLDA*+Matching.

6.2 Metadata Enrichment

The methods used for automatic classification in our experiments were TF-IDF+Jaccard, TF-IDF+Matching, and α-*TaggingLDA*+Matching. For the α-*TaggingLDA*, the corpus builder is based on the search results obtained by querying Googles' search API. The titles and short text summaries (snippets) of the ten most relevant results returned are used to create ten different textual documents. The final *ad hoc* corpus for the learning object consists of the former and the textual content of the resource. Then, by applying LDA (with Gibbs sampling implementation provided by the Machine Learning for Language Toolkit - MALLET[12]) to this corpus we extracted the desired number of latent topics. The default number of topics considered was two, according to the optimal setting specified in [5]. From these topics, the top tags were inferred and matched against the domain topics table presented in Section 5. The method produces a score for each topic in the classification where the top topic (highest score) was chosen to classify the input document.

6.3 Evaluation I: User Classification

The goal of this study was to collect evidence to evaluate if the automatic classification actually matches the categories assigned by the users.

This evaluation is a user study in which each participant was presented with basic information regarding a document, namely, the title and an abstract varying from 60 up to 500 words (see Figure 2). Each document was randomly selected from the dataset. The format of the original resource (e.g. video, image, presentation or document) was not made known to the participants in order to align the nature of the evaluation and to avoid biased judgments of the classification relevance based on non computer-understandable information.

Each participant was then instructed to read the title and the description of the document and finally to choose one of the categories in the proposed domain classification, as depicted in Figure 2. Once submission of the form was completed, the participant was presented with a new object to be evaluated. Additionally, the participants had the option to skip at any point the analysis of a given document, whenever they did not understand the meaning of the content or did not feel confident judging it. We kindly asked each participant to repeat the process for at least ten objects; however, we did not limit their maximum contribution to the study.

In order to assess the quality of the results of this evaluation, we measured the agreement between participants' choices and automatic classifications, and we also used recall, precision and F_1 measure, three widely used metrics. The metrics are defined in Equations 5 and 6.

- Recall for a given classification c is defined as:

$$recall = \frac{|ClassifiedDocs(c) \cap AutoClassifiedDocs(c)|}{|ClassifiedDocs(c)|} \tag{5}$$

- Precision for a given classification c is defined as:

$$precision = \frac{|ClassifiedDocs(c) \cap AutoClassifiedDocs(c)|}{|AutoClassifiedDocs(c)|} \tag{6}$$

where *ClassifiedDocs(c)* is the set of documents assigned to a category c by a participant and *AutoClassifiedDocs(c)* is the set of documents assigned to a category c by the automatic classifier. The aggregated values of recall and precision are then used to compute their harmonic mean or f1 measure as defined according to Equation 7.

$$f1 = 2 \cdot \frac{recall \cdot precision}{recall + precision} \tag{7}$$

6.4 Evaluation II: User Agreement

The goal of this experiment was to evaluate the quality of the automatically assigned categories. Similarly to Evaluation I, in this user study, each participant was presented with the title and an abstract of a randomly selected document. Once again, due to the same reasons presented before, the format of the original resource was not disclosed to participants. In addition, participants were presented with a suggested topic classification (see Table 2 for the list of possible classifications). Note that in this evaluation we only presented to participants classifications given by the proposed α-*TaggingLDA*+Matching method.

[12] `http://mallet.cs.umass.edu`

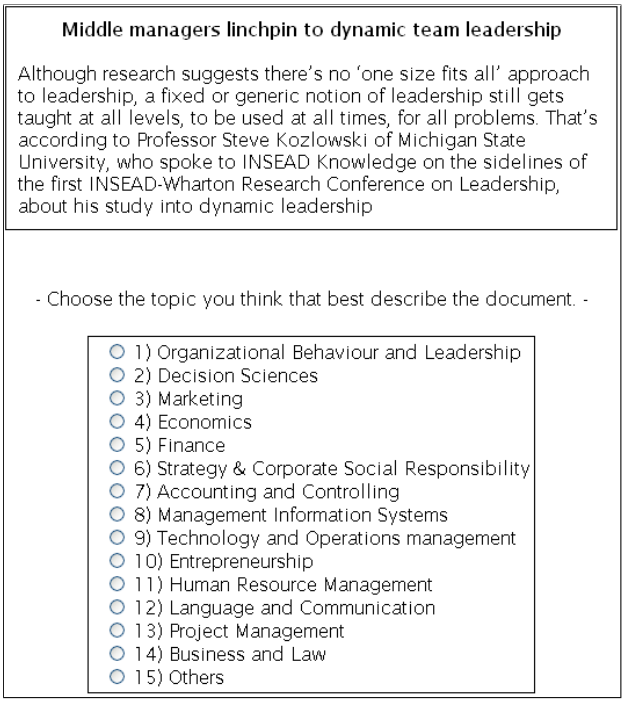

Figure 2: Evaluation I: User Classification Interface. Participants were instructed classify the documents in one of the proposed categories.

Table 3: The overlap of the classifications given by each combination of methods with the classifications given by the participants.

Evaluation 1 - Results		
Participants Classification	658	-
TF-IDF + Jaccard	103	15.7%
TF-IDF + Matching	157	24.0%
α-TaggingLDA + Matching	209	31.8%

Participants were then instructed to read the title and description of the document and finally to rate, in a 5-point Likert scale, their level of agreement with the proposed classification (Figure 3). Once submission of the form was completed, participants were presented with a new object to be evaluated. Once again, participants were provided with an option to skip to a next document in case they did not understand the meaning of the content or did not feel confident judging it.

6.5 Participants' Behavior Analysis

In addition to the 5-point Likert scale in Evaluation II, and the user classification in Evaluation I, in the background of both evaluations we actively logged the participants' behavior during the tasks. In order to evaluate the degree of difficulty of the classification, we logged how long each participant took for analyzing each document in each evaluation and how many times they skipped a given document.

7. RESULTS

Table 4: Precision, Recall and f1-score for each strategy.

Strategy	Precision	Recall	f1
TF-IDF + Jaccard	0.30	0.22	0.20
TF-IDF + Matching	0.26	0.26	0.25
α-TaggingLDA + Matching	0.37	0.35	0.33

Table 5: The results of the participants agreement with the automatic classification given by the α-TaggingLDA+Matching.

Evaluation 2 - Results	
Strongly disagree	9%
Disagree	12%
Neither agree or disagree	7%
Agree	40%
Strongly agree	32%

In our user study, we had a total of 81 participants (31 female and 50 male); 51 of them explicitly stated to be students and 18 were professionals in the area of education. Their average age was 32, ranging from 19 to 66 years old. In total, participants evaluated 658 documents (405 unique) during the first part of the evaluation and 765 (478 unique) items during the second.

With the data collected in the first part of the evaluation, we compared the participants' categorization with those automatically assigned by the different methods (Table 3). The

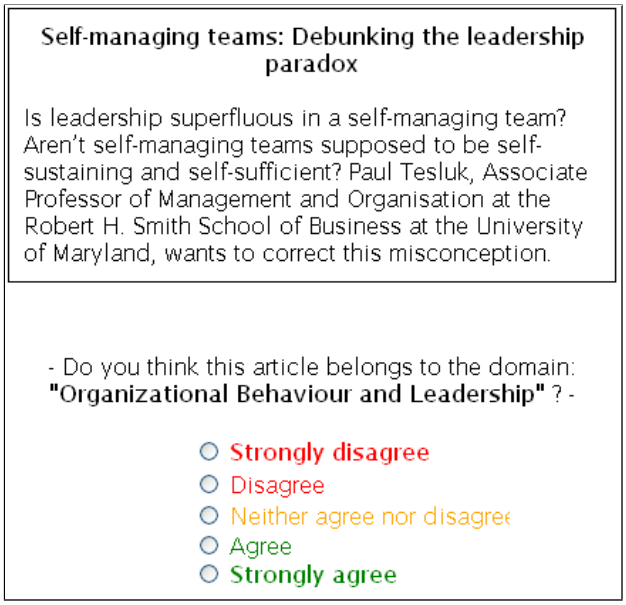

Figure 3: Evaluation II: User Agreement Interface. Participants were instructed to read the abstract of the document and decide (in a 5-point Likert Scale) their agreement with suggested classification.

best performing method, α-*TaggingLDA*+Matching, produces a 32% improvement over the TF-IDF+Matching method.

The results for recall, precision and f1 are exposed in Table 4. Although the values of recall and precision are not exceptionally high, α-*TaggingLDA*+Matching provides significant improvement over the other strategies.

Additionally, from the feedback during the second stage of our user study we found that in 72% of the cases, participants strongly agreed or agreed with the automatic classification assignments given by α-*TaggingLDA*+Matching (Table 5). These results show that participants are inclined to accept the suggested categorization even though it may not be their first choice. We hypothesize that this happens because the task of classifying a document is more complex than only judging if a category is correct or not.

Since in our evaluation setup each document could be assigned to only one exclusive category, participants were facing the paradox of choice, when given several categories to choose from. To better expose this idea we logged the time participants took to perform each task. During the stage which only required the judgment whether a category was relevant or not, in average the participants took 27.0 seconds per document.

During the category assignment task, the average time was 36.6 seconds (35% higher). Additionally, we computed the number of times participants skipped a certain task. During the judgment stage there were 33 skips while in the other task there were two times more, 65 skips in total.

8. GENERAL APPLICATION

In order to further demonstrate the effectiveness of our method, and its flexibility to cover other general domains, we performed an experiment on a categorized newspaper's articles dataset. The dataset consists of a sample of 21,578 articles from Reuters news agency, between February 1987 to October 1987, provided by UCI KDD Archive[9]. Each document is manually annotated with one or more subcategories that belong to a top category.

In our experiments, we automatically assigned the given article to one of these top categories. To setup our automatic classifier, we employed the subcategories as relevant keywords (around 89 keywords per category) that best represented the top category (there were 5 top categories). Our automatic domain classifier achieved 79% accuracy. In comparison, a bag-of-words strategy, where terms are extracted from the articles instead of automatic tags, achieved 71% accuracy.

9. CONCLUSION

In this work, we empirically demonstrated through a series of evaluations that our automatic classification method produces quality enhancements for unclassified documents. Results of the second evaluation show that, in 72% of the cases, automatic classification was relevant and well accepted by participants, which answers our first question: *"Q1: From the user's perspective, how relevant for the document is the automatic classification?"*. This implicitly shows that automatic classifications assigned to documents can effectively support information retrieval, especially in case of cold-start scenarios.

The outcomes of the first part of the evaluation demonstrate that the agreement between automatic categorization and that given by participants does not achieve impressive numbers. However, given the fact that we are working with a cold-start scenario, we still managed to reach an accuracy of almost one third, a 32% improvement over the baseline. This answers our second proposed question: *"Q2: To what extent the classifications assigned automatically agree with*

those given by human judgments?"

Most important is the assessment of participants' behaviors during the two given tasks. We observed that assigning a category to a document is indeed more difficult and more time-consuming than judging a suggested categorization. The evidence lies on the logs obtained during the evaluations. The same participants took in average 35% more time and refrained from making a decision twice more often during the categorization task.

Finally, the good results achieved in our experiments with Reuters' dataset positively answer our last question *"Q3: Are automatic classifications useful and effective in other domains?"*. We believe that these outcomes clearly support our proposed automatic classifying method, and we hypothesize that this tool can be of utility for suggesting categorization to unassigned documents in digital libraries.

We also imagine that greater results can be obtained when the given categories do not have such a succinct difference, as for example *'economics'* and *'finance'*. Additionally, one could apply α-*TaggingLDA* on top of the plain categories descriptions to obtain a representation of each category through tags that are implicitly given by the wisdom of the crowds. In this sense, we predict better matching results since the terms are not given only by a few domain experts but tend to converge to a common vocabulary that is closer to the final consumer of the library.

10. ACKNOWLEDGEMENT

The described work was funded by the European Commission within the eContentplus targeted project grant ECP 2008 EDU 428016 (OPENSCOUT) and by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement nÂř 270239 (ARCOMEM). Additional thanks to Karina Flosi.

11. REFERENCES

[1] S. Bethard, S. Ghosh, J. H. Martin, and T. Sumner. Topic model methods for automatically identifying out-of-scope resources. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 19–28, NY, USA, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] T. M. Department, T. Minka, and J. Lafferty. Expectation-propagation for the generative aspect model. In *In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359. Morgan Kaufmann, 2002.

[4] E. Diaz-Aviles, M. Fisichella, R. Kawase, W. Nejdl, and A. Stewart. Unsupervised auto-tagging for learning object enrichment. In *EC-TEL*, volume 6964 of *Lecture Notes in Computer Science*, pages 83–96. Springer, 2011.

[5] E. Diaz-Aviles, M. Georgescu, A. Stewart, and W. Nejdl. Lda for on-the-fly auto tagging. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 309–312, New York, NY, USA, 2010. ACM.

[6] I. B. et al. A latent dirichlet model for unsupervised entity resolution. In *SDM*, 2006.

[7] M. Fisichella, A. Stewart, K. Denecke, and W. Nejdl. Unsupervised public health event detection for epidemic intelligence. In J. Huang, N. Koudas, G. Jones, X. Wu, K. Collins-Thompson, and A. An, editors, *CIKM*, pages 1881–1884. ACM, 2010.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004.

[9] S. Hettich and S. D. Bay. The uci kdd archive, 1999.

[10] T. Joachims. Text categorization with support vector machines: Learning with many relevant features, 1998.

[11] T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, MA, USA, 2002.

[12] T.-K. Kim, H. Kim, W. Hwang, and J. Kittler. Component-based lda face description for image retrieval and mpeg-7 standardisation. *Image Vision Comput.*, 23(7):631–642, 2005.

[13] A. Kolcz and W. tau Yih. Raising the baseline for high-precision text classifiers. In P. Berkhin, R. Caruana, and X. Wu, editors, *KDD*, pages 400–409. ACM, 2007.

[14] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 524–531. IEEE Computer Society, 2005.

[15] A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In S. McDonald and J. Tait, editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 181–196. Springer, 2004.

[16] K. Niemann, U. Schwertel, M. Kalz, A. Mikroyannidis, M. Fisichella, M. Friedrich, M. Dicerto, K.-H. Ha, P. Holtkamp, and R. Kawase. Skill-based scouting of open management content. In *EC-TEL*, volume 6383 of *Lecture Notes in Computer Science*, pages 632–637. Springer, 2010.

[17] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 81–90, New York, NY, USA, 2010. ACM.

[18] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47, 2002.

[19] P. Soucy and G. W. Mineau. Beyond tfidf weighting for text categorization in the vector space model. In L. P. Kaelbling and A. Saffiotti, editors, *IJCAI*, pages 1130–1135. Professional Book Center, 2005.

[20] S. Veeramachaneni, D. Sona, and P. Avesani. Hierarchical dirichlet model for document classification. In L. D. Raedt and S. Wrobel, editors, *ICML*, volume 119 of *ACM International Conference Proceeding Series*, pages 928–935. ACM, 2005.

[21] D. Xing and M. Girolami. Employing latent dirichlet allocation for fraud detection in telecommunications. *Pattern Recogn. Lett.*, 28(13):1727–1734, 2007.