JIS

Article

Towards Linking Libraries and Wikipedia: Automatic Subject Indexing of Library Records with Wikipedia Concepts

Journal of Information Science
1–11

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Arash Joorabchi
Department of Electronic and Computer Engineering, University of Limerick, Ireland

Abdulhussain E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Ireland

Abstract

In this article, we first argue the importance and timely need of linking libraries and Wikipedia for improving the quality of their services to information consumers, as such linkage will enrich the quality of Wikipedia articles and at the same time increase the visibility of library resources which are currently overlooked to a large degree. We then describe the development of an automatic system for subject indexing of library metadata records with Wikipedia concepts as an important step towards Library-Wikipedia integration. The proposed system is based on first identifying all Wikipedia concepts occurring in the metadata elements of library records. This is then followed by training and deploying generic machine learning algorithms to automatically select those concepts which most accurately reflect the core subjects of the library materials whose records are being indexed. We have assessed the performance of the developed system using standard information retrieval measures of precision, recall, and F-score on a dataset consisting of 100 library metadata records manually indexed with a total of 469 Wikipedia concepts. The evaluation results show that the developed system is capable of achieving an averaged F-score as high as 0.92.

Keywords

Text mining; metadata generation; subject metadata; library metadata; bibliographic records; Wikipedia

1. Introduction

Over the last decade, libraries have experienced a steady decline in the number of users of their online catalogues and websites, which in turn translates into a substantial decrease in the number of information consumers such as students and scholars who use library resources for their information needs. Reportedly, very few information searches (~1%) start from library websites, while the great majority of the rest of information seeking activities (84%) start from search engines such as Google (62%) [1]. On the other hand, Google-Wikipedia is becoming a prevalent information-seeking paradigm in which the information seeker submits an informational query, i.e., query on a particular topic, subject, or concept, to Google and then follows one of the search results redirecting to a relevant article on Wikipedia. Reportedly, Wikipedia appears on page one of the Google search results for 60% of informational queries and in 66% of such cases it appears in top-visibility positions (1-3) of the results page, where majority of clicks occur [2].

Wikipedia is the world's largest web-based free-content encyclopedia project. The English Wikipedia alone currently contains more than four million articles [3]. Wikipedia articles are written, edited, and kept up-to-date and accurate (to a large degree) by a vast community of volunteer contributors, editors, and administrators who are collectively called Wikipedians. An investigation conducted by Nature in 2005 [4] suggested that Wikipedia comes close to Encyclopaedia Britannica in terms of the accuracy of its science entries, which was later disputed by the Britannica [5]. However, regardless of occasional controversies around the accuracy of its articles, Wikipedia is serving a significant role in

Corresponding author:
Abdulhussain E. Mahdi, Department of Electronic and Computer Engineering, University of Limerick, Limerick, Republic of Ireland.
Email: Hussain.Mahdi@ul.ie

fulfilling public information needs. For example, results of a nationwide survey conducted in the U.S. in 2007 showed that Wikipedia attracted six times more traffic than the next closest website in the 'educational and reference' category and preceded websites such as Google Scholar and Google Books with a large margin [6].

In the context described above, linking Wikipedia articles to the records of relevant library materials will give information seekers the option to readily acquire lists of library resources which provide them with more in-depth knowledge on their subject of interest. In this paradigm each Wikipedia article will be linked to the records of relevant materials in a global union catalogue of libraries around the world, WorldCat.org, which in turn provides bibliographic information on the materials of interest and directs information seekers to their local libraries, where they can access those materials. Introduction of this new Wikipedia-Library information seeking paradigm will consequently improve the visibility of library resources which are currently overlooked to a large extend by those information consumers with lower information literacy skills. Looking at it from another perspective, in this paradigm Wikipedia plays the role of a new controlled vocabulary for subject indexing of library materials as an alternative (and/or compliment) to the traditional controlled vocabularies currently used in libraries, such as the Library of Congress Subject Headings (LCSH). As a controlled vocabulary, Wikipedia offers a number of unique advantages over traditional controlled vocabularies:

(1) Extensive coverage and comprehensiveness: the English Wikipedia currently contains over 4 million articles covering subjects in all aspects of human knowledge and growing. Whereas, the latest version of LCSH (33rd edition), which is the de facto standard controlled vocabulary used in libraries around the world, contains approximately a total of 337,000 subject headings and references [7].
(2) Up-to-dateness: due to the crowd-sourced nature of Wikipedia and its large pool of editors, Wikipedia articles are generally well-maintained and kept quite up-to-date. For example, a recent study examining the potential of combining Twitter and Wikipedia data for event detection shows that in case of major events Wikipedia lags Twitter only by about three hours [8].
(3) Rich description: Wikipedia articles provide rich descriptive content for the represented concepts. Whereas, traditional library controlled vocabularies offer very little or no description for their subject headings. For example, the LCSH authority record for the subject heading 'Metadata'[1] offers only two pieces of information about this subject: (a) the subject may also be referred to by 'data about data' and 'meta-data'; (b) the subject is related to two more specific subjects 'Dublin Core' and 'Preservation metadata'. In contrast, the Wikipedia article for the concept of 'metadata'[2] provides a rich description of the concept including definition, variations, and applications complemented with links and references to relevant materials and related concepts. Therefore, extending the metadata records of library materials with Wikipedia concepts gives the users of online library catalogues the option to find descriptive information about unfamiliar indexing subjects that they may encounter as they browse and explore the collection.
(4) Semantic richness: like the LC subject headings, each Wikipedia article has a descriptor which is the preferred and most commonly used term for the represented concept and it is also assigned a set of non-descriptors which are the less common synonyms and alternative lexical forms for the concept. Also, similar to the concept of 'Related Terms' in library controlled vocabularies, in Wikipedia, related articles are connected via hyperlinks. Furthermore, each Wikipedia article is classified according to the Wikipedia's own community-built classification scheme into one or more broader categories which resembles the concept of 'Broader Terms/Narrower Terms' used in the library controlled vocabularies. Finally, similar to the library controlled vocabularies, Wikipedia addresses the problem of word-sense ambiguity by allowing an ambiguous term to correspond to multiple concepts each representing a different sense of the term, e.g., Java (programming language), Java (town), Java (band), etc.
(5) Multilingual: as of September 2013 Wikipedia exists in more than 285 languages. Wikipedia has more than one million articles in each of the 8 most populated languages and more than one hundred thousand articles in each of the 38 less populated languages [9]. This high level of multilingualism in Wikipedia would allow the design and development of effective multilingual information retrieval systems for libraries once their metadata records are indexed with Wikipedia concepts.

According to a report published by OCLC (Online Computer Library Centre) in 2009 [10], the majority of end users of online library catalogues surveyed considered adding more subject information to the metadata records of library materials to be the most helpful enhancement in relation to improving the discovery-related data quality of the catalogues. Using Wikipedia as a complementary controlled vocabulary in library catalogues addresses this need by

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

offering a vast pool of fine-grained subject headings to complement the traditional library controlled vocabularies currently used for subject indexing of library materials.

Based on above, subject indexing of library records with Wikipedia concepts may be considered the first step towards a full Wikipedia-Library integration. However, the high cost of manual subject indexing for libraries combined with the explosive growth in the number of new published materials poses a major obstacle to achieving this goal. Therefore, considering libraries' limited resources, our aim is to reduce the cost of such integration as much as possible. To this end, in this article we describe the design and development of a new software system for automatic subject indexing of library records with Wikipedia concepts. There has been substantial research carried out in relation to automating the process of subject indexing of library records with traditional library controlled vocabularies and classification systems; Golub [11] and Yi [12] provide detailed reviews of such works. However, to the best of our knowledge, the current work is the first attempt at automatic subject indexing of library records with Wikipedia concepts.

The rest of the article is organized as follows: Section 2 lays out our vision for a full Wikipedia-Library integration. Section 3 describes the proposed automatic subject indexing system and its implementation details. Section 4 describes the evaluation process and presents its results. This is followed by Section 5 which provides a conclusion along with a summary account of planned future work.

2. Wikipedia-Library Integration

In our vision, a full Wikipedia-Library integration would create a bi-directional link and flow of information and users between the Wikipedia and libraries. In such environment, information seekers may start their search activities from either of these sources and traverse back and forth as needed.

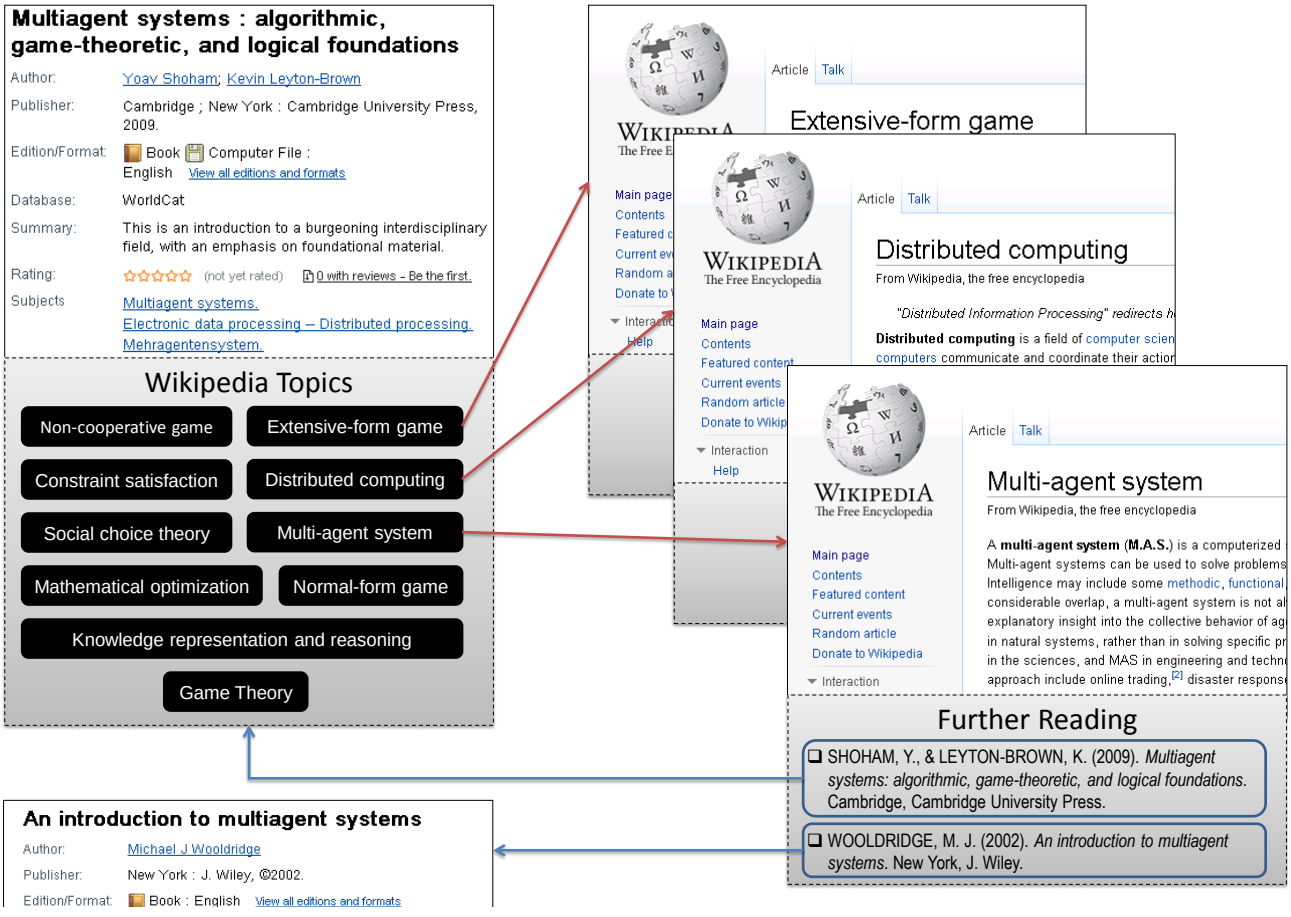

Figure 1. proposed Wikipedia-Library intelinkage

As depicted in Figure 1, on the library side, users who are searching and browsing a library's catalogue would be able to see the subject metadata of each item in form of a set of Wikipedia topics which are hyperlinked to their corresponding Wikipedia articles. Annotation of library items with such complementary subject metadata not only allows users to search and browse library collections via Wikipedia topics, but also offers users the option to find detailed information regarding those topics on the Wikipedia site when encountering unfamiliar topics. On the Wikipedia side however, once a user is on the Wikipedia page of a topic, either after redirection from a library catalogue or via other means, such as conducting a Google or Wikipedia search, he/she will be provided with a "further reading" list. This reading list consists of a set of library resources most relevant to that particular topic. Each item in the list will link to its corresponding record on WorldCat.org, which in turn enables the user to check the availability of the item in his/hers local library(ies). We believe developing such interlinkage among library records and Wikipedia articles, and the subsequent full-circle flow of information and users between the Wikipedia and libraries, would greatly help these organizations in their shared primary goal to effectively assist their users in their information seeking activities.

Fulfilling the proposed vision of Wikipedia-Library integration requires two major developments: (a) annotation of library records with Wikipedia topics as subject metadata; and (b) annotation of Wikipedia articles with citations to the most relevant library resources. In Sections 3 & 4, we describe the design, development, and testing of an automated system to address the first challenge. In the last section of the article, we discuss the requirements of the second challenge and provide an outline of our proposed solutions as future work.

3. Methodology

Our proposed approach to automatic subject indexing of library records with Wikipedia concepts comprises two main stages: (a) identifying Wikipedia concepts appearing in the content of the metadata record to be indexed; and (b) binary classification of detected concepts into key or non-key concepts. We have utilized an open-source toolkit called Wikipedia-Miner [13] for detecting Wikipedia concepts occurring in the content of library metadata records to be indexed. Wikipedia-Miner effectively unlocks the Wikipedia as a general-purpose knowledge source for natural language processing (NLP) applications by providing rich semantic information on concepts and their lexical representations. We use the topic detection functionality of the Wikipedia-Miner to identify all the Wikipedia concepts (i.e., Wikipedia articles) whose descriptor or non-descriptor lexical representations occur in records. After identifying all the Wikipedia concepts occurring in a metadata record, the next step is to distinguish those concepts which are key in terms of reflecting the core subject(s) of the item represented by the record, and are suitable to be added as subject metadata to the record. This distinction is made based on a set of fifteen statistical, positional, and semantical features devised to capture various characteristic of those candidates which have the highest keyness probability:

(1) **Position:** processing the content of library metadata records, we search three specific metadata fields for candidate Wikipedia concepts, namely: "Subject Headings", "Name, and "Description" (ordered according to significance). The occurrence positions of detected concepts are encoded using a three-digit binary number, where the first significant digit represents existence or non-existence of the concept in the least significant metadata field, i.e., "Description", similarly the second and third digits correspond to the "Name" and "Subject Headings" fields, respectively. The resulting binary number is then converted to a decimal number in [1, 7], where, for example, a position value of 1 means that the concept has only occurred in the "Name" field, whereas a position value of 7 means that the concept has appeared in all three fields.

(2) **Frequency:** the occurrence frequency of the candidate concept (i.e., descriptor of the concept) and its synonyms and alternative lexical forms/near-synonyms (i.e., non-descriptors of the concept) in the record. The Frequency values are normalized by dividing them by the highest Frequency value in the record.

(3) **Length:** the number of words in the descriptor of the candidate concept. This feature reflects the general observation that multi-word phrases have a higher keyness probability as they tend to be more specific and less ambiguous. The keyphrase annotation studies which adopt this feature (e.g., see [14, 15]) compute the length of a candidate phrase by simply counting its number of words or characters. However, our approach is to: (a) split the hyphenated words, (b) count the stopwords as 0.5 and non-stopwords as 1.0, (c) normalize the count value by dividing it by 10.0, (d) eliminate candidates which either have a normalized length value greater than 1.0 or those which do not contain any letters (e.g., numbers, numerical dates). Using this weighting scheme reduces some of the noise introduced to the length feature by stopwords. For example, the phrase "de-hyphenation" would count as 1.5 words since "de" is a stopword, and its normalized length value would be 0.15. Eliminating candidates with normalized length values of greater than 1.0 restricts valid candidates to those containing a

maximum of 10 non-stopwords or combinations of stop and non-stop words with a length (measured using the weighting scheme described above) not exceeding 10.0. This is a rather high value for maximum length compared to that adopted by similar works, which usually do not include candidate phrases longer than 3-5 words (counting stopwords as equal as non-stopwords). However, analysis of a dump of the English Wikipedia from July 2011 [16], shows that for a total of 3,573,789 topics, 522,512 (14.6%) have a length (measured using our weighting scheme) in range of 3.5-5.0, 155,220 (4.3%) have a length in range of 5.5-10.0, and 4,083 (0.1%) have a length in range of 10.5 up to a maximum of 32.0. Based on this observation we decided to include all the candidates with a length value up to 10.0 and, hence, excluded only 0.1% of potential candidates, which are highly unlikely to be picked by human indexers.

(4) **Lexical Diversity:** the descriptor and/or non-descriptors of a candidate concept could appear in a record in various lexical forms. We calculate the lexical diversity by (a) case-folding and stemming all the lexical forms of the candidate concept which appear in the record, using an improved version of Porter stemmer [17] called the English (Porter2) stemming algorithm [18]; (b) counting the number of unique stems minus one, so that the lexical diversity value would be zero if there is only one unique stem.

(5) **Average Link Probability:** the average value of the link probabilities of all the candidate concept's lexical forms which appear in the record. The link probability of a lexical form is the ratio of the number of times it occurs in Wikipedia articles as a hyperlink (directing to its corresponding article) to the number of times it occurs as plain text.

(6) **Max Link Probability:** the maximum value of all link probabilities of the lexical forms for a candidate concept which appear in the record. Both the average and max link probability features are based on the assumption that candidate concepts whose descriptor and/or non-descriptor lexical forms appearing in the record have a high probability of being used as a hyperlink in Wikipedia articles, would also have a high keyness probability in metadata records.

(7) **Average Disambiguation Confidence:** in many cases a term in a record could correspond to multiple concepts in Wikipedia and hence needs to be disambiguated. For example, the term "Java" could refer to various concepts, such as "Java programming language", "Java Island", "Java coffee", etc. As described in [19], the Wikipedia-Miner uses a novel machine learning-based approach for word-sense disambiguation which yields an F-measure of 97%. We have set the disambiguator to perform a strict disambiguation, i.e., each term in a record can only correspond to a single concept which has the highest probabilistic confidence. The value of the *average disambiguation confidence* feature for a candidate concept is calculated by averaging the disambiguation confidence values of its descriptor and non-descriptor lexical forms that appear in the record.

(8) **Max Disambiguation Confidence:** the maximum disambiguation confidence value among the lexical forms of a candidate concept which appear in the record. Both the average and max disambiguation confidence features are incorporated to reduce the keyness likelihood of those candidate concepts which have a low disambiguation confidence. A low disambiguation confidence value for a candidate concept sheds doubt on its existence and validity in the record.

(9) **Link-Based Relatedness to Other Concepts:** the Wikipedia-Miner measures the semantic relatedness between concepts using a new approach called Wikipedia Link-based Measure (WLM). In this approach the relatedness between two Wikipedia articles/concepts is measured according to the number of Wikipedia concepts which discuss/mention and have hyperlinks to both the two concepts being compared (see [20] for details). For example, "text mining" and "genetic algorithms" have 53% relatedness based on the fact that a third Wikipedia concept "artificial intelligence" has mentioned and have hyperlinks to both. The *link-based relatedness to other concepts* feature value of a candidate is calculated by measuring and averaging its relatedness to all the other candidates in the record.

(10) **Link-Based Relatedness to Context:** the only difference between this feature and the *link-based relatedness to other concepts* is that the relatedness of the candidate concept is only measured against those of other candidate concepts in the record which are unambiguous, i.e., their descriptor and/or non-descriptor lexical forms occurring in the record have only one valid sense. Both the *link-based relatedness to context* and *link-based relatedness to other concepts* features are incorporated to increase the likelihood of those candidate concepts with high semantic relevance to other concepts in the record being picked as key concepts. However, the former only takes into account the unambiguous concepts in the record and therefore has high accuracy but low coverage, whereas the latter also includes the ambiguous concepts which have been disambiguated based on their surrounding unambiguous context (i.e., unambiguous concepts in the record) and therefore has lower accuracy but conclusive coverage.

(11) **Category-Based Relatedness to Other Concepts:** since May 2004, wikipedians have been categorizing Wikipedia articles according to a community-built classification scheme (a.k.a folksonomy) which has been growing rapidly. The English Wikipedia dump from July 2011, which has been used in this work, contains a total of 739,980 unique categories. This shows 809% growth since January 2006 when it was reported to contain only 91,205 categories [21]. However, in contrast to traditionally expert-built library classification schemes and taxonomies, such as Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) which adhere to a hierarchical tree structure, the Wikipedia classification scheme has a loose semi-hierarchical directed-graph structure which allows articles to belong to multiple categories and categories to have multiple parents. The collaborative and crowdsourcing nature of taxonomy development and categorization work in Wikipedia makes it prone to some level of noise. For example, our analysis of the Wikipedia dump used in this study and those done by others (e.g., see [22]) have shown the existence of self-loops ($C_1 \rightarrow C_1$), direct-loops ($C_1 \rightarrow C_2 \rightarrow C_1$), and indirect-loops (e.g., $C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_1$) among some categories in Wikipedia classification graph. Our study shows that as of July 2011, 95% of Wikipedia articles are classified and on average each classified article belongs to 3.82 categories. When a candidate concept is classified, we can utilize its categorization data to measure its semantic relatedness to other candidates in the record. One of the well-known approaches to estimate the relatedness between two concepts in a taxonomy is to measure the distance of the shortest path between the two nodes in terms of the number of edges along the path, first proposed by Rada et al. [23] in 1989. An enhanced version of this approach, which counts the number of nodes instead of edges along the shortest path and normalizes the resulting distance by dividing it by two times the maximum depth of the taxonomy (as the longest possible distance), was proposed in 1998 by Leacock and Chodorow [24], and used to measure the relatedness between two terms in WordNet as:

$$\text{Relatedness}(term_1, term_2) = -\log \frac{\text{Distance}(term_1, term_2)}{2 \times \text{maximum depth of taxonomy}} \quad (1)$$

In 2006, Strube and Ponzetto [21] adopted this measure to estimate the semantic relatedness between two concepts in Wikipedia and showed its superiority compared to other measures proposed in the literature up to then. In 2008, Milne and Witten [20] showed that their Wikipedia Link-based Measure (WLM), implemented in Wikipedia-Miner and utilized in this work (features 9 and 10), outperforms the shortest-path measure. Nevertheless, we believe deploying these two approaches together would improve the overall performance of our system, as they estimate the semantic relatedness of concepts very differently using two independent information sources in Wikipedia and therefore could complement each other. We measure the category-based relatedness of two Wikipedia concepts as:

$$\text{Relatedness}(concept_1, concept_2) = 1 - \frac{\text{Distance}(concept_1, concept_2) - 1}{2D - 3} \quad (2)$$

where D is the maximum depth of the taxonomy, i.e., 16 in case of the Wikipedia dump used in this work. The distance function returns the length of the shortest path between $concept_1$ and $concept_2$ in terms of the number of nodes along the path. The term $2D-3$ gives the longest possible path distance between two concepts in the taxonomy, which is used as the normalization factor, i.e., $2\times16-3=29$. The shortest possible distance between two nodes/concepts is 1 (in case of siblings) and the longest is $2D-3$. Therefore subtracting one from the outcome of the distance function results in a highest possible relatedness value of 1.0, e.g., $1-(1-1)/(2\times16-3)=1.0$, and a lowest possible relatedness value of 0.03, e.g., $1-(29-1)/(2\times16-3)=0.03$. Changing the divisor from $2D-3$ to $2D-4$ reduces the lowest possible relatedness value to zero, however we have adopted the former and instead assign a zero value to relatedness when either $concept_1$ or $concept_2$ are amongst the 5% of Wikipedia concepts which are not classified. We have used an open-source toolkit for graph modelling, analysis, and visualization called JUNG [25], to build the classification graphs of the records and measure the shortest path distance between the candidate concepts. The value for *category-based relatedness to other concepts* for each candidate is calculated by measuring and averaging its category-based relatedness to all the other candidates in the record.

(12) **Generality:** the depth of the candidate concept in the taxonomy measured as its distance from the root category in Wikipedia, normalized by dividing it by the maximum possible depth, and inversed by deducting the normalized value from 1.0. It ranges between 0.0 for the concept farthest from the root and unclassified ones, and 1.0 for the root itself.
(13) **In Links:** total number of distinct Wikipedia concepts which are linked in to the candidate concept.
(14) **Out Links:** total number of distinct Wikipedia concepts which are linked out from the candidate concept.
(15) **Translations Count:** number of languages that the candidate concept is translated to in the Wikipedia. This feature reflects the assumption that candidate concepts which have been translated to more languages in Wikipedia would have a higher significance and keyness probability.

After identifying all the candidate concepts in a record and computing their feature values, the next step is to detect key concepts based on their feature values. We have approached this problem as a supervised binary classification problem, where each candidate is classified as either key or non-key by a generic machine learning algorithm. Using this approach, a set of library metadata records manually indexed with key Wikipedia concepts is used as training data to learn a model for the key concepts based on their feature characteristics. The learnt model is then applied to an unlabelled set of records used as test dataset to classify the identified candidate concepts in the records as key or non-key and measure the prediction accuracy of the model. As described in Section 4, we have experimented with a host of generic machine learning algorithms and have also evaluated the effectiveness of the above features using various feature selection metrics.

4. Experimental Results & Evaluation

In order to evaluate the accuracy performance of the proposed subject indexing system, a collection of library metadata records manually subject indexed with Wikipedia concepts was required to train and test the system. However, to the best of our knowledge no such dataset has been built before and, therefore, we had to build our own. For this purpose we decided to manually index a small subset of WorldCat-Million dataset with Wikipedia concepts to train and test our system on. The WorldCat-Million dataset released by the OCLC (Online Computer Library Centre) in 2012 [26] contains metadata records of nearly 1.2 million library materials most widely held in libraries around the world. We built our WorldCat-Wikipedia dataset by randomly selecting 100 records from the WorldCat-Million dataset which belonged to the DDC class 006.3, Artificial Intelligence[3]. We limited the sample records to those in this particular DDC class because of our familiarity with its subject and respective terminology and concepts, as such knowledge is essential for the required manual subject indexing task. We used the Wikipedia-Miner to identify the candidate concepts in the 100 selected records and then computed their feature values as described in Section 3. A total of 1,762 candidate concepts were identified in these records with a minimum of 2, average of 17, and maximum of 69 concepts per record. We then examined all the candidate concepts manually and labelled them as key or non-key in respect to their corresponding records. This resulted in a total of 469 candidates being labelled as key and the remaining 1,293 of them as non-key concepts. In total, 26% of identified candidate concepts have been labelled as key and on average each record in the dataset has been indexed by 4.7 key Wikipedia concepts. Table 1 shows a sample record from the dataset. As can be seen in this table, the two FAST subject headings [27] assigned to the record, namely "Multiagent systems" and "Electronic data processing -- Distributed processing", have also been identified as key Wikipedia concepts "Multi-agent system" and "Distributed computing" suitable for indexing the record. Furthermore, these two main concepts have been complemented with 8 more key Wikipedia concepts which reflect more detailed aspects of the work which were not captured by the manually assigned traditional library subject headings.

After manual classification of all the candidate concepts identified in the records of the dataset, the concepts and their corresponding feature values were stored in an Attribute-Relation File Format (ARFF) to be imported into the Weka environment [28]. Weka is an open-source software with a comprehensive collection of machine learning algorithms for data mining tasks. We have used Weka to experiment with various machine learning-based classification algorithms for automatic classification of candidate concepts identified in library metadata records as key or non-key.

Table 2 shows the performance of various well-known generic classification algorithms which we have experimented with and their corresponding performance on the WorldCat-Wikipedia dataset measured using standard information retrieval metrics and 10-fold cross-validation.

Table 1. Sample library metadata record from the WorldCat-Wikipedia dataset.

Record No.	43
URL	http://www.worldcat.org/oclc/213408653
Title	Multiagent systems : algorithmic, game-theoretic, and logical foundations
Description	Distributed constraint satisfaction -- Distributed optimization -- Introduction to noncooperative game theory: games in normal form -- Computing solution concepts of normal-form games -- Games with sequential actions: reasoning and computing with the extensive form -- Richer representations: beyond the normal and extensive forms -- Learning and teaching -- Communication -- Aggregating preferences: social choice -- Protocols for strategic agents: mechanism design -- Protocols for multiagent resource allocation: auctions -- Teams of selfish agents: an introduction to coalitional game theory -- Logics of knowledge and relief -- Beyond belief: probability, dynamics, and intention., This is an introduction to a burgeoning interdisciplinary field, with an emphasis on foundational material.
FAST Subject Headings	Multiagent systems, Electronic data processing -- Distributed processing
Key Wikipedia Concepts	Multi-agent system, Distributed computing, Game theory, Mathematical optimization, Knowledge representation and reasoning, Normal-form game, Extensive-form game, Social choice theory, Non-cooperative game, Constraint satisfaction.
Non-Key Wikipedia Concepts	Computer, Education, Logic, Electronics, Communication, Algorithm, Knowledge, Probability, Theory, Data, Solution, Design, Sequence, Learning, Communications protocol, Auction, Reason, Field (mathematics), Strategy, Belief, Economic system, Material, Selfishness, Surface normal, Relief, Dynamical system, Foundations of mathematics, Resource, Interdisciplinarity, Computer data processing, Swarm behaviour, Preference, Computing, Mechanism design, Social, Data (computing), Agent (economics), Resource allocation, Constraint (mathematics), Contentment, The Normal, Fifth normal form.

Table 2. Classification performance of various classification algorithms on the WorldCat-Wikipedia dataset

Classifier (Weka implementation)	TP Rate	FP Rate	Precision	Recall	F_1	MCC	ROC Area	PRC Area
KNN (IB1 instance-based classifier)	0.865	0.225	0.864	0.865	0.864	0.651	0.820	0.823
SVM (LibSVM)	0.890	0.283	0.898	0.890	0.882	0.711	0.804	0.827
Decision Tree (J48)	0.892	0.174	0.892	0.892	0.892	0.723	0.875	0.863
Bayes Network (BayesNet)	0.899	0.174	0.898	0.899	0.898	0.738	0.947	0.955
Random Forest (RandomForest)	0.907	0.144	0.908	0.907	0.908	0.763	0.947	0.946
Bagging Random Forest	0.916	0.155	0.915	0.916	0.915	0.781	0.956	0.959
Random Committee Random Forest	0.920	0.136	0.919	0.920	**0.920**	0.793	0.960	0.964
Random Committee Random Forest + Feature Selection	0.921	0.144	0.920	0.921	0.920	0.793	0.960	0.963

As shown in Table 2, Random Forest family of classifiers consistently yielded higher performance than other tested classifiers, with the Random Committee Random Forest classifier achieving the best performance of 0.92 F_1.

We applied four commonly used feature selection metrics, namely Chi-squared, Info Gain, Correlation, and Symmetrical Uncertainty, which were implemented in Weka, to the WorldCat-Wikipedia dataset to determine the effectiveness of each of the 15 features defined for the candidate concepts in Section 3. Figure 2 shows the average normalized ranks for each feature according to above four feature selections metrics after 10-fold cross-validation. According to presented results the 2nd feature (F2), Frequency, has achieved the lowest rank amongst other features and therefore may be regarded as the weakest feature with the lowest positive impact on the accuracy performance of the classification algorithms. We tested this assumption by re-training the best preforming classification algorithm on the dataset, i.e. Random Committee Random Forest, this time excluding the F2. The last row of Table 2 shows the results of this test and confirms that F2 does not help learning a more accurate classification model from the data. This may be attributed to the fact that majority of library metadata records have limited textual content and, therefore, the

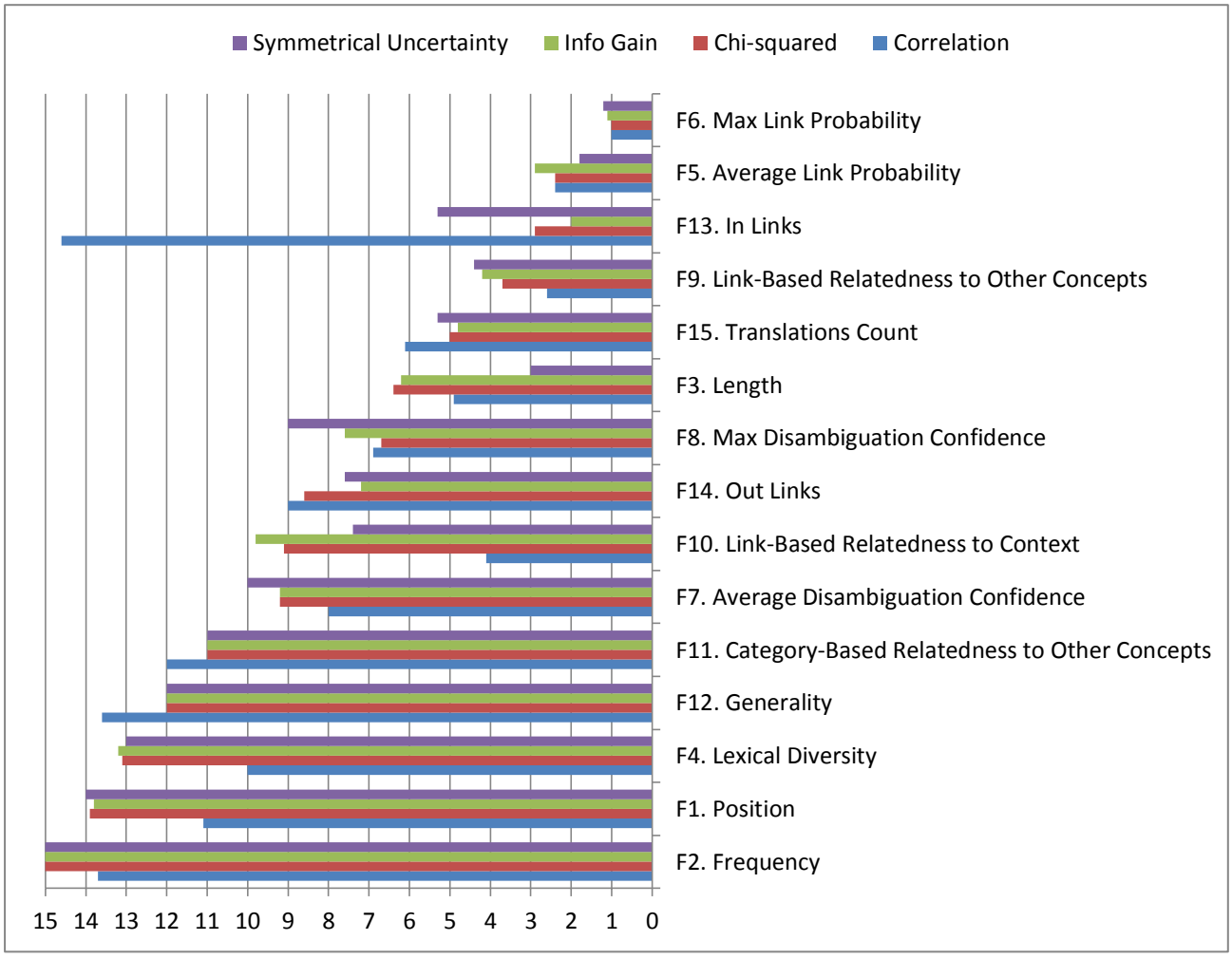

Figure 2. Average normalized ranks for each future according to four feature selections metrics

reoccurrence probability of candidate concepts is quite low. Consequently this feature is rendered obsolete. We did the same experiment with the next lowest ranking feature, F1, which encodes the occurrence positions of candidate concepts in records. However, eliminating F1 from the feature set proved to have a considerable adverse effect on the classification performance. Repeating the same experiment with the rest of features showed that, apart from F2, all other features are necessary for achieving the best possible performance.

Looking at the other end of spectrum, F6 and F5, both measuring the link probability of candidate concepts, have shown to be strongest features. This means that the number of times a concept has occurred as a hyperlinked term in the content of Wikipedia articles has a strong correlation with the probability of that concept being a suitable candidate for subject indexing library metadata records. Also, expectedly, the rankings of features F13 and F14 show that the number of times that a concept has been linked to from within the content of other Wikipedia concepts/articles has a higher significance than the number of times that links are made from its content to other concepts.

As discussed and predicated in Section 3, features F9 and F10, which measure the relatedness of a candidate concept to other concepts in the record based on the WLM approach, have outranked F11 which does the same using the shortest-path approach. Finally, the relatively high ranks achieved by F15 and F3 show that the number of languages in which a given concept has been represented in Wikipedia and the length of a concept measured as proposed in Section 3 can be used as strong indicators regarding the keyness of candidate concepts detected in records to be indexed.

All the data related to above experiments is available for download[4]. This includes: (a) the metadata records retrieved from the WorldCat-Million dataset which belong to the DDC class 006.3, “Artificial Intelligence” in JSON format; (b) a

log file containing the data produced during the process of detecting candidate concepts in the records and computing their feature values; and (c) the manually annotated WorldCat-Wikipedia dataset in ARFF format which may be easily used to duplicate all the reported experiments and conduct further experiments on in the Weka environment.

5. Conclusion and Future Work

In this article, we discussed the benefits of automatic subject indexing of library metadata records with Wikipedia concepts as a first step towards Library-Wikipedia integration; and described the design and development of a machine learning-based system capable of automating the proposed indexing process with a high level of accuracy. The encouraging result of this study paves the way for development of robust automatic subject indexing systems in libraries and also future research towards realizing the ultimate goal of full Library-Wikipedia integration. In this context, there are two paths for further research and development: (a) further development of our prototype into a fully-fledged automatic subject indexing software tool which may be readily integrated into library workflows independently or through services such as WorldCat.org, WorldCat Local, and WorldShare Management Services provided by the OCLC[29]; (b) automatic annotation of Wikipedia articles with most relevant library resources: this represents the next necessary step towards full Library-Wikipedia integration, in which each Wikipedia article will be automatically annotated with a list of library resources which are most suitable as further reading for those who are seeking more information on the concept represented by the article. This task involves semantic comparison of the textual content of each Wikipedia concept (article) with that of those library metadata records which have been indexed with that concept to find the best matches. The result of this semantic matching process may be then extended by taking into account other parameters such as the popularity of best semantically matching resources in terms of the number of libraries which hold them, their up-to-dateness, and ratings on LibraryThing, Goodreads and other similar social cataloguing web applications. Taking into account all these factors, the proposed system should then be able to compile a compelling list of library resources for further readings to be added to the Wikipedia article being annotated. We envisage the actual task of annotation to be also automated using a Wikipedia bot. Wikipedia bots are software systems capable of editing Wikipedia automatically. The big advantage of using bots is their ability to analyse and edit Wikipedia articles on a large scale without any human intervention and in a time efficient manner. This advantage of using bots becomes more evident when one considers the size of both the Wikipedia and the WorldCat and their fast growth rate, which make the task of linking them manually extremely labour intensive. Obviously, these bots cannot be expected to be as efficient as humans in terms of the accuracy of the links that they create between individual Wikipedia articles and WorldCat records. However, we believe they can offer an acceptable level of accuracy which may be further improved manually by wikipedians. In effect, these bots will bootstrap and catalyse the process of Wikipedia-WorldCat integration.

Notes

1. http://id.loc.gov/authorities/subjects/sh96000740.html
2. http://en.wikipedia.org/wiki/Metadata
3. http://dewey.info/class/006.3/e23/2012-10-24/about.en
4. http://www.skynet.ie/~arash/zip/WorldCat-Wikipedia.zip

Funding

This research was partly funded under the ‘Research & Practice in ICT Learning’ initiative – University of Limerick.

References

[1] De Rosa C., *Perceptions of libraries and information resources : a report to the OCLC membership* (OCLC Online Computer Library Center, Dublin, Ohio, 2005).

[2] Safran N., *Wikipedia in the SERPs*, 2012, http://www.conductor.com/blog/2012/03/wikipedia-in-the-serps-appears-on-page-1-for-60-of-informational-34-transactional-queries/ (accessed July 2013)

[3] *Wikipedia:Size in volumes*, (Wikimedia Foundation, Inc., 2013), http://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes (accessed July 2013)

[4] Giles J., Internet encyclopaedias go head to head, *Nature* 2005; 438, 7070: 900-901.

[5] *Fatally Flawed - Refuting the recent study on encyclopedic accuracy by the journal Nature*, (Encyclopædia Britannica, Inc., 2006), http://corporate.britannica.com/britannica_nature_response.pdf (accessed July 2013)

[6] Rainie L. and Tancer B., *Wikipedia users*, (Pew Internet and American Life Project, 2007), http://www.pewinternet.org/Reports/2007/Wikipedia-users.aspx (accessed July 2013)

[7] *Library of Congress Subject Headings*, http://www.loc.gov/cds/products/product.php?productID=44 (accessed July 2013)
[8] Osborne M., Petrovic S., McCreadie R., Macdonald C. and Ounis I. Bieber no more: First Story Detection using Twitter and Wikipedia. In: SIGIR Workshop in Time-aware Information Access (TAIA'12); 2012 August 12–16; Portland, Oregon, USA: ACM; 2012.
[9] *List of Wikipedias*, (Wikimedia Foundation, Inc., 2013), http://en.wikipedia.org/wiki/List_of_Wikipedias (accessed August 2013)
[10] Calhoun K., Cantrell J., Gallagher M., Hawk J. and Gauder B., *Online catalogs : what users and librarians want : an OCLC report* (OCLC, Dublin, Ohio, 2009).
[11] Golub K., Automated subject classification of textual Web pages, based on a controlled vocabulary: Challenges and recommendations, *New Review of Hypermedia and Multimedia* 2006; 12, 1: 11-27.
[12] Yi K., Automated Text Classification Using Library Classification Schemes: Trends, Issues, and Challenges, *International Cataloguing and Bibliographic Control (ICBC)* 2007; 36, 4: 78-82.
[13] Milne D. An open-source toolkit for mining Wikipedia. In: New Zealand Computer Science Research Student Conference; 2009; 2009.
[14] Turney P. D., Learning Algorithms for Keyphrase Extraction, *Information Retrieval* 2000; 2, 4: 303-336.
[15] Medelyan O., Witten I. H. and Milne D. Topic Indexing with Wikipedia. In: first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08); 2008; Chicago, US: AAAI Press; 2008.
[16] *enwiki dump progress on 22/07/2011*, (Wikimedia dump service, 2011), http://dumps.wikimedia.org/enwiki/20110722/ (accessed 11 March 2012)
[17] Porter M. F., An algorithm for suffix stripping, *Program* 1980; 14, 3: 130−137.
[18] M.F.Porter, *The English (Porter2) stemming algorithm*, (Snowball, 2002), http://snowball.tartarus.org/algorithms/english/stemmer.html (accessed 11 March 2012)
[19] Milne D. and Witten I. H. Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management; 2008; Napa Valley, California, USA: ACM; 2008. p. 509-518.
[20] Milne D. and Witten I. H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08); 2008; Chicago, I.L; 2008.
[21] Strube M. and Ponzetto S. P. WikiRelate! computing semantic relatedness using wikipedia. In: proceedings of the 21st national conference on Artificial intelligence - Volume 2; 2006; Boston, Massachusetts: AAAI Press; 2006. p. 1419-1424.
[22] Salah A. A., Gao C., Suchecki K. and Scharnhorst A., Need to Categorize: A Comparative Look at the Categories of Universal Decimal Classification System and Wikipedia, *Leonardo* 2012; 45, 1: 84-85.
[23] Rada R., Mili H., Bicknell E. and Blettner M., Development and application of a metric on semantic nets, *Systems, Man and Cybernetics, IEEE Transactions on* 1989; 19, 1: 17-30.
[24] Leacock C. and Chodorow M. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database*. In C. Fellbaum (Ed.), MIT Press, 1998, p. 265-283.
[25] O'Madadhain J., Fisher D., Nelson T., White S. and Boey Y.-B., *JUNG 2.0*, (Released under the open source GPL licence, 2009), http://jung.sourceforge.net/index.html (accessed 11 March 2012)
[26] Murphy B., *OCLC provides downloadable linked data file for the 1 million most widely held works in WorldCat*, (OCLC, DUBLIN, Ohio, 2012), http://www.oclc.org/news/releases/2012/201252.en.html (accessed July 2013)
[27] Dean R. J., FAST: Development of Simplified Headings for Metadata, *Cataloging & Classification Quarterly* 2004; 39, 1-2: 331-352.
[28] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I. H., The WEKA Data Mining Software: An Update, *SIGKDD Explorations* 2009; 11, 1.
[29] *OCLC (Online Computer Library Center*), http://www.oclc.org/ (accessed July 2013)