



## Improving the Visibility of Library resources Via Mapping Library Subject Headings to Wikipedia Articles

Journal:	<i>Library Hi Tech</i>
Manuscript ID	LHT-04-2017-0066.R2
Manuscript Type:	Original Article
Keywords:	Library catalogues, Wikipedia, FAST subject headings, Controlled vocabularies, Semantic mapping, Data integration

SCHOLARONE™  
Manuscripts

Hi Tech

# Improving the Visibility of Library resources Via Mapping Library Subject Headings to Wikipedia Articles

## Abstract

**Purpose** – Linking libraries and Wikipedia can significantly improve the quality of services provided by these two major silos of knowledge. Such linkage would enrich the quality of Wikipedia articles and at the same time increase the visibility of library resources. To this end, this work describes the design and development of a software system for automatic mapping of FAST subject headings, used to index library materials, to their corresponding articles in Wikipedia.

**Design/methodology/approach** – The proposed system works by first detecting all the candidate Wikipedia concepts (articles) occurring in the titles of the books and other library materials which are indexed with a given FAST subject heading. This is then followed by training and deploying a Machine Learning (ML) algorithm designed to automatically identify those concepts that correspond to the FAST heading. The ML algorithm used is a binary classifier which classifies the candidate concepts into either “corresponding” or “non-corresponding” categories. The classifier is trained to learn the characteristics of those candidates which have the highest probability of belonging to the “corresponding” category based on a set of fourteen positional, statistical, and semantic features.

**Findings** – We have assessed the performance of the developed system using standard information retrieval measures of precision, recall, and F-score on a dataset containing 200 FAST subject headings manually mapped to their corresponding Wikipedia articles. The evaluation results show that the developed system is capable of achieving F-scores as high as 0.65 and 0.99 in the corresponding and non-corresponding categories respectively.

**Research limitations/implications** – The size of the dataset used to evaluate the performance of the system is rather small. However, we believe the developed dataset is large enough to demonstrate the feasibility and scalability of the proposed approach.

**Practical implications** – The sheer size of English Wikipedia makes the manual mapping of Wikipedia articles to library subject headings a very labour-intensive and time consuming task. Therefore, our aim is to reduce the cost of such mapping and integration.

**Social Implications** – The proposed mapping paves the way for connecting libraries and Wikipedia as two major silos of knowledge, and enables the bi-directional movement of users between the two.

**Originality/value** – To the best of our knowledge, the current work is the first attempt at automatic mapping of Wikipedia to a library controlled vocabulary.

## Keywords

Library catalogues, Wikipedia; FAST subject headings, Controlled vocabularies, Semantic mapping, Data integration

## 1. Introduction

Library websites and online catalogues are experiencing a decline in their number of visitors. This, in turn, could translate into a decrease in the number of students and other information seekers who use library resources. According to De Rosa (2005), less than 1% of online information searches start from library websites, and the majority of the rest of information seeking activities (~84%) start from search engines such as Google. This wide spread low-effort information seeking behaviour is known by the library and information science scholars as the Principle of Least Effort (PLE) (Chang, 2016). According to this principle the main concern underlying the majority of information seeking behaviours is the desire to reduce the time and effort spent, as formalized by the Zipf's law (Zipf, 1949).

Subsequently, Google-Wikipedia is becoming a prevalent online information seeking route. In this new trend, the information seeker submits an informational query (i.e., query on a particular topic, subject, or concept) to Google and follows one of the search results to a relevant article on Wikipedia. Safran (2012) showed that Wikipedia articles appear on page one of Google search results for 60% of informational queries, and in 66% of such cases Wikipedia articles

1  
2  
3 appear in top-visibility positions (1-3) of the results page, where the majority of clicks occur. A more recent case study  
4 by McMahon et al. (2017) on the relationship between Wikipedia and Google demonstrates an extensive and mutually  
5 beneficial interdependence between the two. In this study Wikipedia links were silently removed from the search results  
6 presented to the participants to examine the effect. Reportedly, the quality of Google search results considerably  
7 degrades for many queries when links to Wikipedia content are excluded; the study also highlights Google's important  
8 role in providing readership to Wikipedia.

9 Wikipedia has become the largest free encyclopedia online. The English Wikipedia currently contains over five  
10 million articles. Wikipedia articles are written and edited by a large community of volunteer contributors, editors, and  
11 administrators. Wikipedia serves an important role in addressing public information needs. For example, results of a  
12 nationwide survey conducted in the U.S in 2007 showed that 36% of American Internet users look for information on  
13 Wikipedia; and Wikipedia attracted six times more traffic than the next closest website in the "educational and  
14 reference" category, outperforming websites such as Google Scholar and Google Books with a large margin (Rainie and  
15 Tancer, 2007). This nationwide survey was repeated again in 2010 and showed that the rate of American Internet users  
16 who turn to Wikipedia for information has risen from 36% to 53%, and Wikipedia is most popular with the 18-29 age  
17 group (Zickuhr and Rainie, 2011).

18 In the context described above, linking Wikipedia articles to the records of related library materials would enable  
19 information seekers to readily acquire lists of library resources which provide in-depth knowledge on their subject of  
20 interest. In this paradigm each Wikipedia article would be linked to the records of related materials in a global union  
21 catalogue of libraries around the world, i.e., WorldCat.org. This in turn would provide bibliographic metadata on the  
22 materials of interest and direct information seekers to their local libraries, where they can access those materials.  
23 Availability of this new Wikipedia-Library information seeking paradigm would consequently improve the visibility of  
24 library resources which are currently overlooked to a large extent by those information consumers with lower  
25 information literacy skills.

26 Based on above, mapping Wikipedia articles to their corresponding library subject headings (i.e., LCSH, FAST)  
27 could play an important role towards Wikipedia-Library integration. In practice, such mapping would enable the bi-  
28 directional movement of users between libraries and Wikipedia as two major silos of knowledge. However, the sheer  
29 size of English Wikipedia (>5m articles) makes the manual mapping of Wikipedia articles to library subject headings a  
30 very labour-intensive and time consuming task. Therefore, our aim is to reduce the cost of such mapping and  
31 integration. To this end, in this article we describe the design and development of a new software system for automatic  
32 mapping of Wikipedia articles to their corresponding FAST subject headings. There has been substantial research  
33 carried out in relation to automating the process of subject indexing of library records and electronic documents with  
34 traditional library controlled vocabularies and classification systems. Golub (2006) and Yi (2007) have reviewed earlier  
35 works in this field carried out by library organizations such as Library of Congress and Online Computer Library Center  
36 (OCLC); and Wang (2009) and Khoo et al. (2015) have reviewed more recent works in the context of metadata  
37 management in digital libraries and repositories. Also, Wikipedia has been successfully used in various information  
38 search and retrieval applications as an intermediary resource to improve indexing and classification, and optimize users'  
39 queries. For example, Hinze et al. (2015) used Wikipedia as a knowledgebase of concepts to create a semantic-enhanced  
40 search service for the HathiTrust Digital Library (HTDL). Joorabchi et al. (2015) proposed an automatic method for  
41 mapping user tags (folksonomy) to their corresponding concepts in Wikipedia. Deveaud et al. (2012) proposed two  
42 query expansion approaches involving Wikipedia as an external source of information for book search. Shapira et al.  
43 (2015) have created a taxonomy describing the application domains of text analytics in which Wikipedia can be utilized  
44 and referenced relevant studies in each domain.

45 However, to the best of our knowledge, the current work is the first attempt to automatically map Wikipedia to a  
46 library controlled vocabulary. In an earlier work (Joorabchi and Mahdi, 2014), we proposed an automatic method for  
47 subject indexing of individual library records with Wikipedia concepts as an initial step towards Library-Wikipedia  
48 integration. The current work enhances and improves on our previous method by moving the linking/mapping process  
49 from the level of individual library records to the higher level of library subject headings. The advantages of linking the  
50 subject headings used to index the records instead of the records themselves are twofold: first, it enables creating a bi-  
51 directional link between the Wikipedia and library catalogues; and second, it eliminates the need for indexing each  
52 newly created library record individually.

53 The rest of the article is organized as follows: Section 2 lays out our vision for a full Wikipedia-Library integration.  
54 Section 3 describes the proposed automatic mapping system and its implementation details. Section 4 describes the  
55 evaluation process and presents its results. This is followed by Section 5 which provides a conclusion along with a  
56 summary of planned future work.  
57  
58  
59  
60

## 2. Benefits of Wikipedia-Library Integration

Wikipedia-Library integration would create a bi-directional link and flow of information and users between the Wikipedia and libraries. This would enable information seekers to start their search activities from either of these sources and traverse back and forth as needed. As illustrated in Figure 1, users on the library side, who are searching and browsing the library's catalogue, would be able to see the subject metadata of each item in the form of a set of FAST subject headings linked to their equivalent Wikipedia articles. Creating such linkage not only allows users to search and browse library collections via Wikipedia topics/concepts, but also enables users to find detailed information on those topics on the Wikipedia when encountering unfamiliar ones. On the Wikipedia side, once a user reaches a Wikipedia article on a topic via conducting a Google or Wikipedia search, he/she would be provided with a link to the topic's equivalent FAST subject heading(s) on the WorldCat.org website. These links would enable Wikipedia users to find and browse all the library resources relevant to a given Wikipedia article and check their availability in their local libraries. We believe developing such interlinkage between library records and Wikipedia articles, and the subsequent bi-directional flow of information and users between the Wikipedia and library catalogues, would greatly serve the shared primary goal of these organizations to effectively assist their users in their information seeking activities.

**Figure 1.** Proposed vision for a full Wikipedia-Library integration.

## 3. Automatic Mapping of FAST Subject Headings to Wikipedia Articles

The FAST subject headings are divided into eight different facets (personal names, corporate names, geographic names, events, titles, time periods, topics, and form/genre) and amount to a total of 1.7 million headings across all facets (Dean, 2004). The initial focus of our project is on mapping the 400,000 topical subject headings (MARC Field 650) to their corresponding Wikipedia articles, as establishing such mapping would be the most fruitful in terms of realising the proposed vision of full library-Wikipedia integration. On the Wikipedia side, the English version of Wikipedia currently contains over 5 million articles whose equivalent FAST headings could belong to any of the 8 facets of FAST. The job of the automatic mapping algorithm is to find the most probable matching Wikipedia articles for the FAST headings. Figure 2 shows an outline of the proposed algorithm. Our proposed method to automatic mapping of FAST subject headings to their corresponding Wikipedia article(s) comprises three main stages:

- (1) Data collection: retrieving titles of library materials indexed with the FAST heading to be mapped.
- (2) Candidate detection: identifying all the candidate Wikipedia concepts appearing in the collected titles.
- (3) Candidate classification: binary classification of detected candidates as either "corresponding" or "non-corresponding".

**Figure 2.** Proposed automatic mapping method.

### 3.1. Data collection

We have used a locally stored version of the FAST dataset[1] in MARCXML format to iterate through the FAST records which are to be mapped, and retrieve their details to be used during the mapping process. As shown in Figure 2, the process starts by retrieving a list of all the books and other library materials which are indexed with the given FAST heading. This is achieved by submitting a REST query to the OCLC Classify API[2] in the following format:

```
http://classify.oclc.org/classify2/Classify?ident=[FAST Control Number]&maxRecs=100&summary=false&orderBy=hold%20desc
```

This query returns the metadata records of up to 100 books which are indexed with the given FAST heading. The returned records are sorted according to their number of library holdings in a descending order. We iterate through all the returned records, extract their titles, and compile them into a single text file. The content of the "titles" text file provides a rich source of keywords related to the FAST heading to be mapped, and therefore it can be used to find the most

probable Wikipedia article(s) for the heading. For example, consider the FAST heading “AIDS (Disease)--Prevention”[3] which is equivalent to the article “Prevention of HIV/AIDS”[4] in Wikipedia. Querying the Classify API for books indexed with this heading returns many titles[5], most of which contain relevant terms and keywords. Figure 3 shows the top 40 matching titles with the relevant terms and keywords highlighted. Hartley (2005) argues: “*Whatever the format, book titles are remarkable for conveying a good deal of information in very few words*”. But, as the example in Figure 3 shows, titles are not always sufficiently descriptive of book content; and, more importantly in our case, they do not always contain relevant terms and keywords. However, our approach relies on a set of titles (rather than a single title), which are all indexed with the same FAST heading, to collectively provide a descriptive set of terms and keywords for the mapping task.

We also enrich the “titles” text file by adding the *FAST heading* (preferred label), *see from tracings* (alternative labels), and *see also headings* (related terms) to the beginning of the file. When available, adding these metadata elements of the FAST record to the “titles” file would increase the chance of finding the right corresponding Wikipedia article(s) for the FAST heading. For example, in case of the FAST heading “Abdomen--Surgery”, the *see from tracings* are “Abdominal surgery” and “Laparotomy”. The title (preferred label) of the corresponding Wikipedia article for this FAST heading is “Abdominal surgery”, and therefore, there is an exact matching between one of the *see from tracings* of the FAST record and the title of its corresponding Wikipedia article.

**Figure 3.** Top 40 titles indexed with the FAST heading “AIDS (Disease)--Prevention”.

### 3.2. Candidate detection

The second stage of the mapping process involves detecting all the candidate Wikipedia articles/concepts appearing in the “titles” file of a given FAST heading. This is achieved using an open-source toolkit called Wikipedia-Miner[6] (Milne and Witten, 2013). Wikipedia-Miner effectively unlocks 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 the “titles” files. The parameters of the topic detector and its disambiguator component are set such that all the possible candidate concepts are detected, regardless of their degree of probability (*disambiguator.setMinSenseProbability(0)*, *disambiguator.setMinLinkProbability(0)*, *topicDetector.setDisambiguationPolicy(DisambiguationPolicy.LOOSE)*, *topicDetector.allowDisambiguations(false)*). This process results in detecting hundreds of candidate Wikipedia concepts in the “titles” file of a FAST heading. For example, a total of 380 Wikipedia concepts were detected in the “titles” file of the FAST heading “AIDS (Disease)--Prevention”, some of which include: Prevention of HIV/AIDS, HIV/AIDS, HIV, Epidemiology of HIV/AIDS, HIV/AIDS in China, HIV/AIDS in the United States, HIV-1, History of HIV/AIDS, HIV/AIDS in South Africa, Discredited HIV/AIDS origins theories, HIV/AIDS denialism, International AIDS Society, HIV/AIDS in Africa, Diagnosis of HIV/AIDS, Circumcision and HIV, HIV-positive people, Preventive healthcare, Human sexual activity, HIV/AIDS research, Drug injection, HIV-associated neurocognitive disorder, Sexually transmitted infection, Management of HIV/AIDS, Transmission (medicine), Vertically transmitted infection, AIDS education and training centers, Epidemic, LGBT, Centers for Disease Control and Prevention, Immunodeficiency, Sex tourism, Immune system, Sexual ethics, HIV/AIDS denialism in South Africa, President's Emergency Plan for AIDS Relief, Immunology, AIDS orphan, Criminal transmission of HIV, Needle sharing, Compulsory sterilization, Reproductive rights, Society, Government, Virus, Outreach.

A considerable number of detected Wikipedia concepts would, to various degrees, be related to the FAST heading. However, only one or a few of them directly correspond to the FAST heading. In case of the above example, the only true corresponding candidate Wikipedia concept is “Prevention of HIV/AIDS”. The mapping relationship between a FAST heading and Wikipedia articles could be of a one-to-one type (e.g., AIDS (Disease)--Prevention → Prevention of HIV/AIDS) or of a one-to-many type. An example of the latter type is the FAST heading “Abnormalities, Human--Genetic aspects” which has two corresponding articles in Wikipedia: “Congenital disorder” and “Genetic disorder”.

### 3.3. Candidate classification

The third stage of the mapping process involves finding the most probable corresponding Wikipedia article(s) for the FAST heading among the large set of candidates detected in the “titles” file. This is achieved using a Machine Learning (ML) based binary classifier which classifies each candidate concept as either “corresponding” or “non-corresponding”. To build and train such a classifier we need to: (a) devise a set of distinguishing features for Wikipedia concepts which help capturing various characteristics of those candidates that have the highest correspondence probability; and (b) manually map a set of sample FAST headings to their corresponding Wikipedia concepts/articles to train the classifier with and evaluate its prediction performance.

#### 3.3.1 Features for candidate Wikipedia concepts

In order for an ML-based classifier to identify a FAST heading’s corresponding Wikipedia concept(s) among all the candidate concepts detected in the heading’s “titles” file, a set of features capturing the properties of the concepts which belong to the “corresponding” category is required. We have devised a set of fourteen positional, statistical, and semantic features to capture various characteristics of those candidates which have the highest probability of belonging to the “corresponding” category:

- (1) 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 FAST heading’s “titles” file. The frequency values are normalized by dividing by the highest frequency value in the candidates set. We expect the FAST heading’s corresponding concept(s) to have a relatively higher occurrence frequency compared to other candidate concepts identified in the “titles” file. The effectiveness of this feature is well proven in similar information retrieval and text mining applications such as automatic topic indexing (Medelyan, 2009) and keyword extraction (Hulth, 2004).
- (2) FAST Record Position: as described in 3.1, the first three lines of the “titles” file contain three metadata fields from the FAST heading’s record, namely, the *FAST heading* (preferred label), *see from tracings* (alternative labels), and *see also headings* (related terms). Therefore, the candidate concepts which appear in any of these first three lines have a significantly higher probability of belonging to the “corresponding” category. The *FAST Record Position* captures and encodes this characteristic of candidate concepts using a three-digit binary number as follows. The least significant bit of this binary number represents the least significant metadata field, i.e., *see also headings*, with 1 if the candidate concept exists in this field and with 0 otherwise. Similarly, values will be allocated to the next significant and most significant digits of the binary number to represent the *see from tracings* and *FAST heading* metadata fields, respectively. The resulting binary number is then converted to decimal, where, for example, a decimal value of 1 means that the concept has only occurred in the *see also headings* field, whereas a value of 7 means that the concept has appeared in all three fields, as illustrated in Table 1.

**Table 1.** Numeric values for the FAST Record Position.

- (3) Lexical Diversity: the descriptor and/or non-descriptors of a candidate concept could appear in a “titles” file 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 file using an improved version of Porter stemmer called the English (Porter2) stemming algorithm (M.F.Porter, 2002); and (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.
- (4) Average Link Probability: the average value of the link probabilities of all the lexical forms of the candidate concept which appear in the “titles” file. The link probability of a lexical form is the ratio of the number of times it occurs in Wikipedia articles as a hyperlink to the number of times it occurs as plain text.
- (5) Max Link Probability: the maximum value of the link probabilities of all the lexical forms of the candidate concept which appear in the “titles” file. Both the average and max link probability features are based on the assumption that the candidate concepts whose descriptor and/or non-descriptor lexical forms appearing in the “titles” file have a high probability of being used as a hyperlink in Wikipedia articles, would also have a higher probability of belong to the “corresponding” category.
- (6) Average Disambiguation Confidence: in many cases a term in a FAST heading’s “titles” file 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 (Milne and Witten, 2008b), the Wikipedia-Miner uses a novel machine learning-based approach for word-sense disambiguation which yields an F-measure of 97%. In this approach the sense of an ambiguous term which corresponds to more than one concept is decided by inferring the main sense of the document (i.e., “titles” file here) as a whole. For example, if the majority of the unambiguous terms in the “titles” file are related to “computer programming”, then it may be inferred that the term “Java” in this context corresponds to the concept “Java programming language”. Developing a semantic-enhanced search method for digital libraries, Hinze et al. (2015) used the same disambiguation method to translate between user keywords/phrases and their respective Wikipedia concepts. As described in Section 3.2, we have set the Wikipedia-Miner’s disambiguator component to perform a loose disambiguation, i.e., each term in the “titles” file could correspond to multiple concepts with various levels of probability. The value of this 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 FAST heading’s “titles” file. This feature acts as a validity check mechanism for the candidate concepts.

- (7) Max Disambiguation Confidence: the maximum disambiguation confidence value among the lexical forms of a candidate concept which appear in the FAST heading’s “titles” file. Both the average and max disambiguation confidence features are incorporated to reduce the “correspondence” likelihood score of those candidate concepts which have a low disambiguation confidence. A low disambiguation confidence value for a candidate concept reduces its validity and questions its existence.
- (8) Link-Based Relatedness to Other Concepts: the Wikipedia-Miner measures the semantic relatedness between concepts using a method called Wikipedia Link-based Measure (WLM). In this method the relatedness between two Wikipedia articles/concepts is measured according to the number of Wikipedia concepts which mention and have hyperlinks to both concepts being compared, see (Milne and Witten, 2008a) for details. For example, “text mining” and “genetic algorithms” have 53% relatedness based on the fact that a third Wikipedia concept “artificial intelligence” mentions and contains hyperlinks to both. The value of this feature for a candidate concept is obtained by measuring and averaging its relatedness to all the other candidates detected in the FAST Heading’s “titles” file. The FAST heading’s corresponding Wikipedia concept(s) is expected to have a high semantic relatedness to the majority of other candidate concepts detected in the heading’s “titles” file, as together they form a cluster of related concepts each covering a specific aspect of the same subject/topic discussed in the heading’s “titles” file.
- (9) 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 “titles” file which are unambiguous, i.e., their descriptor and/or non-descriptor lexical forms occurring in the “titles” file 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 “correspondence” likelihood score of those candidate concepts which have a high semantic relevance to other concepts in the “titles” file. However, the former only takes into account the unambiguous concepts in the wiki page and therefore has a 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 “titles” file) and therefore has a lower accuracy but conclusive coverage.
- (10) 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). The English Wikipedia dump from October 2015, which has been used in this work, contains a total of 4,799,116 articles and 1,321,139 categories. This shows a 14-fold growth in the number of categories since January 2006 when it was reported to contain only 91,205 categories (Strube and Ponzetto, 2006). Our study shows that as of October 2015, 4,658,682 of Wikipedia articles (97.1%) are classified and on average each classified article belongs to 4.42 categories. When a candidate concept is classified, we can utilize its categorization data to measure its semantic relatedness to other candidates in the “titles” file. 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 (Rada et al., 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 by Leacock and Chodorow (1998), and used to measure the relatedness between two terms in WordNet as:

$$\text{Relatedness}(\text{term}_1, \text{term}_2) = -\log \frac{\text{Distance}(\text{term}_1, \text{term}_2)}{2 \times \text{maximum depth of taxonomy}} \quad (1)$$

Strube and Ponzetto (2006) adopted above 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. Milne and Witten (2008a) showed that their Wikipedia Link-based Measure (WLM), implemented in Wikipedia-Miner and utilized in this work (features 8 and 9), 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 would complement each other. We measure the category-based relatedness of two Wikipedia concepts as:

$$\text{Relatedness}(\text{concept}_1, \text{concept}_2) = 1 - \frac{\text{Distance}(\text{concept}_1, \text{concept}_2) - 1}{2D - 3} \quad (2)$$

where  $D$  is the maximum depth of the taxonomy, i.e., 18 in case of the Wikipedia dump used in this work. The distance function returns the length of the shortest path between  $\text{concept}_1$  and  $\text{concept}_2$  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 ( $2 \times 18 - 3 = 33$ ). 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 ( $1 - (1 - 1) / (2 \times 18 - 3) = 1.0$ ), and a lowest possible relatedness value of 0.03 ( $1 - (33 - 1) / (2 \times 18 - 3) = 0.03$ ). 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  $\text{concept}_1$  or  $\text{concept}_2$  are amongst the 2.9% of Wikipedia concepts which are not classified. We have used an open-source toolkit for graph modelling, analysis, and visualization called JUNG (O'Madadhain et al., 2009), 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 FAST heading's "titles" file.

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

### 3.3.2 Building a training & testing dataset

Having defined a set of features for the Wikipedia concepts detected in the "titles" files of the FAST headings, we then need to build a dataset of manually mapped *FAST Headings-to-Wikipedia Concepts* instances. This dataset is fed to an ML-based classification algorithm for learning a prediction model. We also use the same dataset for evaluating the prediction accuracy performance of the classifier using a 10-fold cross-validation procedure (more on that in Section 4). The dataset was built by manually mapping a set of 200 randomly chosen FAST headings to their equivalent Wikipedia concepts/articles. The FAST headings to Wikipedia Articles mapping could be of either a one-to-one type (e.g., Abdomen--Wounds and injuries → Abdominal trauma) or a one-to-many type. The one-to-many mappings occur when:

- (1) There exist multiple Wikipedia articles which correspond to a given FAST heading. For example, the Wikipedia articles "Abortion debate" and "Religion and abortion" both correspond to the FAST heading "Abortion--Moral and ethical aspects".
- (4) The FAST heading is too specific to be mapped to a single Wikipedia article. For example, consider the case of the FAST heading "Aboriginal Australian literature", where, as of the time of this work, there exists no Wikipedia article which focuses specifically on the topic of aboriginal Australian literature. Therefore, alternatively, the heading is mapped to three different Wikipedia articles "Indigenous Australians", "Aboriginal Australians", and "Australian literature", which collectively correspond to the heading by each covering a particular aspect of it.



1  
2  
3 Accordingly, 110 of the total 200 sample headings, which were manually examined, were mapped to their single  
4 corresponding Wikipedia articles (i.e., one-to-one mappings), 60 were mapped to multiple articles (i.e., one-to-many  
5 mappings), and for the remaining 30 headings no corresponding Wikipedia articles were found. Most of the FAST  
6 headings, which we could not find a corresponding article for, represented either out-of-date or very specific concepts  
7 and were used to index only a handful of library materials in the WorldCat catalogue. Some examples of these headings  
8 induce: “AN/BSY-2 (Computer system), WorldCat usage: 2”, “Abashev culture, WorldCat usage: 10”, “ASCOP  
9 (Electronic computer system), WorldCat usage: 1”, and “AAAD Basketball Tournament, WorldCat usage: 1”.

10 Excluding the null mapping cases, the final dataset contains a total of 170 FAST headings which are manually  
11 mapped to 241 Wikipedia articles. The 60 FAST headings, which are in the one-to-many mapping group, are mapped to  
12 a total of 131 articles (i.e., an average of 2.2 articles per heading). Also, the dataset contains a total of 45,181 Wikipedia  
13 articles/concepts which are detected in the FAST headings’ “titles” files as candidates but belong to the “non-  
14 corresponding” category. This means that each FAST heading in the dataset is assigned an average of 267 candidate  
15 Wikipedia articles out of which only 1.1 belong to the “corresponding” category and the rest belong to the “non-  
16 corresponding”. Therefore, building the dataset involved manual verification of a large number of candidate Wikipedia  
17 articles (267 on average) per each FAST heading in the dataset. This proved to be a very time-consuming task and the  
18 main obstacle which prohibited us from building a larger dataset. However, we believe the size of the current dataset is  
19 large enough to demonstrate the feasibility of our proposed approach. Table 2 shows a number of sample mappings  
20 from the dataset.  
21

22 **Table 2.** Sample FAST to Wikipedia mappings.  
23

## 24 4. Experimental Results & Evaluation

25  
26 The dataset described in 3.3.2 is stored in Attribute-Relation File Format (ARFF)[7], which is the main file format used  
27 in Weka environment (Hall et al., 2009). Weka is an open-source data-mining software tool issued under the GNU  
28 General Public License, which offers a comprehensive collection of data mining and machine learning algorithms. We  
29 have used Weka to experiment with and evaluate the accuracy performance of our proposed mapping method which  
30 uses an ML-based binomial classifier at its core. Table 3 shows the evaluation results of our experiments with various  
31 well-known ML-based classification algorithms, measured using standard information retrieval metrics and 10-fold  
32 cross-validation.  
33

34  
35 **Table 3.** Classification performance achieved using various classification algorithms in Weka.  
36

37  
38 As can be seen in the results presented in Table 3, the four classification algorithms used have yielded the same  
39 overall accuracy performance in terms of the weighted average  $F_1$  measure (0.996) and the  $F_1$  measure achieved for the  
40 “non-corresponding” category of instances (0.998). However, the Multilayer Perceptron has outperformed the other  
41 classifiers in terms of the  $F_1$  measure achieved for the “corresponding” category (0.647 vs. 0.606). The precision and  
42 recall performance achieved by all the classifiers for the “non-corresponding” category is close to optimal and for the  
43 “corresponding” category is within an acceptable range. The precision and recall achieved for the “corresponding”  
44 category by the best performing classifier, Multilayer Perceptron, are 0.735 and 0.577 respectively. This shows that  
45 there is still room for improvement in terms of the accuracy achieved in identifying the concepts which belong to the  
46 “corresponding” category. The lower accuracy achieved for the “corresponding” category could be attributed to the fact  
47 that the high number of candidate concepts detected per FAST heading (hundreds) makes identifying the single (in most  
48 cases) true corresponding concepts a non-trivial task. This situation may be improved by either changing the Wikipedia-  
49 Miner parameters (described in Section 3.2) to perform a strict disambiguation and hence reduce the number of detected  
50 candidates in the FAST heading’s “titles” file, or add a new filtering step which would eliminate the candidates whose  
51 values for some features are below a certain threshold (e.g., *frequency* < 2). However, both of these strategies could  
52 increase the precision achieved in the “corresponding” category at the expense of a lower recall rate. As with any other  
53 classification problem, the key issue here is to strike a good balance between the precision and recall achieved by the  
54 classifier. Also, the precedence of one measure over the other could dictate the choice of the classification algorithm  
55 used. For example, as shown in Table 3, the Multilayer Perceptron classifier has achieved the highest recall for the  
56 “corresponding” category (0.577), whereas, the Random Forest has achieved the highest precision in the same category  
57 (0.844) but a lower recall (0.473).  
58  
59  
60

As described in Section 3.1, the content of a FAST heading's "titles" file come from two sources: (1) the heading's metadata, i.e., the *FAST heading* (preferred label), *see from tracings* (alternative labels), and *see also headings* (related terms); and (2) titles of up to 100 books indexed with the given heading. In order to establish the importance of each of these sources and their contribution to the final mapping task, we inspected the source of all the 241 candidate Wikipedia concepts which truly correspond to one of the FAST headings in the dataset (true positives), i.e., all the candidate concepts manually classified as "corresponding". This was achieved by analysing the values of the "FAST Record Position" feature (defined in Section 3.3.1) for all the concepts classified as "corresponding". As the results of this analysis in Tables 4 show, 20 of the corresponding concepts have not appeared in any of their FAST headings' metadata fields. This means that if our method solely relied on its first source (i.e., headings' metadata) for data collection and ignored the second source (i.e., titles of the books indexed with the headings), it would have missed these 20 corresponding concepts. Subsequently, the recall of the system would have dropped by at least 8.3%. In addition to enriching the pool of candidate concepts (i.e., improving recall), using the titles of the books as a second source provides rich textual content for the WikipediaMiner to disambiguate the candidate concepts and measure their relatedness to the context and other candidate concepts with a reliable confidence level. This could improve the quality of features 6-10 defined in Section 3.3.1, and hence result in improving the precision of the system.

The analysis results presented in Table 4 show that the FAST heading's *preferred label* is the most useful metadata element for our mapping task, as in 48% of instances the corresponding concept has appeared in this field. Also, in 42% of cases the corresponding concept has appeared in both the heading's *preferred label* and *alternative labels*. However, there are no instances where the corresponding heading has appeared in the *See also headings* (related terms) field. After inspecting the dataset, we identified this to be due to the fact that values for this metadata field are rarely supplied.

**Table 4.** "FAST Record Position" feature values of the "corresponding" concepts in the dataset.

After evaluating the performance of various classification algorithms for our mapping task, we then used various feature selection metrics to measure the effectiveness of each of the 14 features defined for the Wikipedia concepts in Section 3.3.1. For this purpose, we adopted three commonly-used feature selection metrics, namely Chi-squared, Info Gain, and Correlation, which are all implemented in Weka. Figure 4 shows, in descending order, the average ranks for each feature according to the above three feature selections metrics after 10-fold cross-validation.

As shown in Figure 4, the 10th feature (F10), Category-Based Relatedness to Other Concepts, has achieved the lowest rank among other features, and therefore may be regarded as the weakest feature with the lowest or no positive impact on the accuracy performance of the classification algorithms we have experimented with. We examined this assumption by re-training and testing the best performing classification algorithm (i.e., Multilayer Perceptron) on the dataset, but this time excluding the F10 feature. The last row of Table 3 presents the results of this test which shows, despite its lower rank, excluding F10 has a negative impact on the overall classification performance.

The low rank of F10 may be attributed to the fact that it shows a bias towards more generic Wikipedia concepts detected in the FAST heading's "titles" files. Also, as speculated in Section 3.3.1, the higher ranking of F8 as opposed to that of F9 confirms that the *Link-Based Relatedness to Other Concepts* is a more reliable feature than the *Link-Based Relatedness to Context* for our binomial classification-based mapping task in this work. Looking at the other end of spectrum, F3 and F1 are the first and second high ranking features. F1 captures the occurrence frequency of a candidate concept in a heading's "titles" file and, hence, was expected to act as a strong feature for identifying the concepts which belong to the "corresponding" category. The F3 captures the number of surface forms by which a Wikipedia concept is expressed in the "titles" file. For example, the concept "Prevention of HIV/AIDS" could take any of the following surface forms: "AIDS education", "AIDS prevention", "HIV prevention", "Prevention of AIDS", and "Prevention of HIV". Based on the results of the feature selection, having a candidate concept to appear in the "titles" file in various surface forms (i.e., large F3 value) is the most reliable feature for distinguishing the concepts which belong to the "corresponding" category.

The Last column of Table 3 shows the time taken to build a model using each classifier on a PC platform running Ubuntu 16.04 (64-bit) with an Intel Core 2 Duo Processor E8600 (3.33 GHz x 2) and 16GB of RAM. As these results show, the best performing classifier (i.e., Multilayer Perceptron) has taken 78 seconds; this time has been improved by 9 seconds after eliminating F10, which was identified as the lowest ranking feature in the feature selection stage. Although the dataset used in this work is relatively small, we believe these times provide a good indication of the system's computational demand and potential scalability.

All the data used and generated in this work is available for download[8]. This includes: (a) an excel file containing all the FAST headings and their corresponding Wikipedia articles; (b) a log file containing the data produced during the process of detecting candidate Wikipedia concepts in FAST headings' "titles" files and computing their feature values;

(c) the manually verified *FAST Headings-to-Wikipedia Concepts* mapping dataset in ARFF format, which may be readily used to duplicate all the reported experiments using Weka and to conduct further experimentation and analysis.

**Figure 4.** Average ranks of the Wikipedia concept features according to three different feature selection metrics.

## 5. Conclusion and Future Work

In this work we have described the design and development of an ML-based method for mapping FAST subject headings to their equivalent articles in Wikipedia. The proposed mapping paves the way for connecting libraries and Wikipedia as two major silos of knowledge, and enables the bi-directional movement of users between the two.

In the proposed mapping method, we first detect all the Wikipedia concepts appearing in the titles of the books which are indexed with a given FAST heading. We then deploy an ML-based classification algorithm to classify the detected concepts into “corresponding” and “non-corresponding” categories. We showcased the application of the proposed mapping method and evaluated its performance using a dataset of 170 FAST subject headings manually mapped to their equivalent Wikipedia articles. We evaluated the performance of our method using the standard information retrieval metrics of precision, recall, and  $F_1$ . Depending on the ML-based classification algorithm used,  $F_1$  scores as high as 0.65 and 0.99 were achieved for the “corresponding” and “non-corresponding” categories respectively. Given the non-trivial nature of the mapping task, we believe these results are encouraging. This belief is backed by the evidence that, during the process of manually creating the mapping dataset (described in 3.3.2), we came across a considerable number of one-to-many mapping cases which were difficult for the human annotators to agree on. For example, consider the case of the FAST heading “Aboriginal Australians–Anthropometry” which is mapped to 4 different Wikipedia articles: “Indigenous Australians”, “Aboriginal Australians”, “Biological anthropology”, and “Anthropometry”. In this case, one annotator may consider the mapping to the article “Indigenous Australians” redundant as there already exists a mapping to the article “Aboriginal Australians”, whereas another annotator may consider both mappings necessary. We tried to address this issue by adopting an inclusive mapping strategy and including the disputed articles. Regardless of such issues, the reported results show that there is still room for improving the accuracy of the proposed method. In specific, this may be achieved by optimizing the Wikipedia-Miner to reduce the number of candidate concepts per FAST heading (as discussed in Section 4). Another area for potential improvement is in the data collection stage (Section 3.1), where besides using titles, we may use other metadata elements of books such as their “table of contents” when available.

Analysing the textual content of 4,799,116 Wikipedia articles from the English Wikipedia dump (generated in October 2015) used in this study, showed that a considerable number of articles (375,138) contain at least one valid ISBN number in their “References” section, as shown in Table 5. These ISBNs represent the books which are related to the subjects of the articles, and are cited as further reading sources on the articles’ subjects. Using citation analysis, we could leverage these links between Wikipedia articles and library resources (when available) to enhance our proposed mapping method and potentially improve its accuracy.

**Table 5.** Number of Wikipedia articles citing valid ISBNs.

Also as future work, we plan to demonstrate the application of the proposed mapping method by developing a browser plugin capable of redirecting users, where appropriate, from Wikipedia articles to WorldCat.org website for further reading on their subjects of interest. The plugin would detect if the user is browsing a Wikipedia article, retrieve the article’s corresponding FAST heading(s) from a remote server, and give the user the option to search the WorldCat.org catalogue for the library materials indexed with those FAST headings via a single click. For example, a user looking at the Wikipedia article “Prevention of HIV/AIDS” will be presented with a link[9] to the WorldCat.org which would list all library materials indexed with the subject heading “AIDS (Disease)—Prevention”. The user may then check the availability of any of the listed materials in his/her local library via the WorldCat website. We believe such a service would further showcase our vision of a full Wikipedia-library integration and its benefits, as laid out in Section 3.

## Notes

1. <http://www.oclc.org/research/themes/data-science/fast/download.html>

2. [http://classify.oclc.org/classify2/api\\_docs/index.html](http://classify.oclc.org/classify2/api_docs/index.html)
3. <http://experimental.worldcat.org/fast/793908>
4. [https://en.wikipedia.org/wiki/Prevention\\_of\\_HIV/AIDS](https://en.wikipedia.org/wiki/Prevention_of_HIV/AIDS)
5. <http://classify.oclc.org/classify2/Classify?ident=793908&summary=false&maxRecs=100>
6. <https://github.com/dnmilne/wikipediaminer>
7. <http://www.cs.waikato.ac.nz/ml/weka/arff.html>
8. <http://www.skynet.ie/~arash/zip/FAST2WP.zip>
9. <http://www.worldcat.org/search?q=ar:ocolc%2dfst00793908&qt=searchfast>

## Acknowledgements

This work was supported by the OCLC/ALISE Library & Information Science Research Grant Program (LISRGP) 2016.

## References


- CHANG, Y.-W. 2016. Influence of human behavior and the principle of least effort on library and information science research. *Information Processing & Management*, 52, 658-669.
- DE ROSA, C. 2005. *Perceptions of libraries and information resources : a report to the OCLC membership*, Dublin, Ohio, OCLC Online Computer Library Center.
- DEAN, R. J. 2004. FAST: Development of Simplified Headings for Metadata. *Cataloging & Classification Quarterly*, 39, 331-352.
- DEVEAUD, R., SANJUAN, E. & BELLOT, P. 2012. Social Recommendation and External Resources for Book Search. In: GEVA, S., KAMPS, J. & SCHENKEL, R. (eds.) *Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011, Saarbrücken, Germany, December 12-14, 2011, Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- GOLUB, K. 2006. Automated subject classification of textual Web pages, based on a controlled vocabulary: Challenges and recommendations. *New Review of Hypermedia and Multimedia*, 12, 11-27.
- HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11.
- HARTLEY, J. 2005. To Attract or to Inform: What are Titles for? *Journal of Technical Writing and Communication*, 35, 203-213.
- HINZE, A., TAUBE-SCHOCK, C., BAINBRIDGE, D., MATAMUA, R. & DOWNIE, J. S. 2015. Improving Access to Large-scale Digital Libraries Through Semantic-enhanced Search and Disambiguation. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. Knoxville, Tennessee, USA: ACM.
- HULTH, A. 2004. *Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction*. PhD thesis Ph.D, Stockholm University.
- JOORABCHI, A., ENGLISH, M. & MAHDI, A. E. 2015. Automatic mapping of user tags to Wikipedia concepts: The case of a Q&A website – StackOverflow. *Journal of Information Science*, 41, 570-583.
- JOORABCHI, A. & MAHDI, A. E. 2014. Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts. *Journal of Information Science*, 40, 211-221.
- KHOO, M. J., AHN, J.-W., BINDING, C., JONES, H. J., LIN, X., MASSAM, D. & TUDHOPE, D. 2015. Augmenting Dublin Core digital library metadata with Dewey Decimal Classification. *Journal of Documentation*, 71, 976-998.
- LEACOCK, C. & CHODOROW, M. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database*. In C. Fellbaum (Ed.), MIT Press.

- 1  
2  
3 M.F.PORTER. 2002. *The English (Porter2) stemming algorithm* [Online]. Snowball. Available:  
4 <http://snowball.tartarus.org/algorithms/english/stemmer.html> [Accessed 11 March 2012].  
5  
6 MCMAHON, C., JOHNSON, I. & HECHT, B. 2017. *The Substantial Interdependence of Wikipedia and Google: A*  
7 *Case Study on the Relationship Between Peer Production Communities and Information Technologies*.  
8  
9 MEDELYAN, O. 2009. *Human-competitive automatic topic indexing*. PhD thesis Ph.D, University of Waikato, New  
10 Zealand.  
11  
12 MILNE, D. & WITTEN, I. H. 2008a. An effective, low-cost measure of semantic relatedness obtained from Wikipedia  
13 links. *first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*. Chicago, I.L.  
14  
15 MILNE, D. & WITTEN, I. H. 2008b. Learning to link with wikipedia. *Proceedings of the 17th ACM conference on*  
16 *Information and knowledge management*. Napa Valley, California, USA: ACM.  
17  
18 MILNE, D. & WITTEN, I. H. 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, 194, 222-239.  
19  
20 O'MADADHAIN, J., FISHER, D., NELSON, T., WHITE, S. & BOEY, Y.-B. 2009. *JUNG 2.0* [Online]. Released  
21 under the open source GPL licence. Available: <http://jung.sourceforge.net/index.html> [Accessed 11 March 2012].  
22  
23 RADA, R., MILI, H., BICKNELL, E. & BLETNER, M. 1989. Development and application of a metric on semantic  
24 nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19, 17-30.  
25  
26 RAINIE, L. & TANCER, B. 2007. *Wikipedia users* [Online]. Pew Internet and American Life Project. Available:  
27 <http://www.pewinternet.org/Reports/2007/Wikipedia-users.aspx> [Accessed July 2014].  
28  
29 SAFRAN, N. 2012. *Wikipedia in the SERPs* [Online]. Available: [http://www.conductor.com/blog/2012/03/wikipedia-in-](http://www.conductor.com/blog/2012/03/wikipedia-in-the-serps-appears-on-page-1-for-60-of-informational-34-transactional-queries/)  
30 [the-serps-appears-on-page-1-for-60-of-informational-34-transactional-queries/](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].  
31  
32 SHAPIRA, B., OFEK, N. & MAKARENKO, V. 2015. Exploiting Wikipedia for Information Retrieval Tasks.  
33 *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information*  
34 *Retrieval*. Santiago, Chile: ACM.  
35  
36 STRUBE, M. & PONZETTO, S. P. 2006. WikiRelate! computing semantic relatedness using wikipedia. *proceedings of*  
37 *the 21st national conference on Artificial intelligence - Volume 2*. Boston, Massachusetts: AAAI Press.  
38  
39 WANG, J. 2009. An extensive study on automated Dewey Decimal Classification. *Journal of the American Society for*  
40 *Information Science and Technology*, 60, 2269-2286.  
41  
42 YI, K. 2007. Automated Text Classification Using Library Classification Schemes: Trends, Issues, and Challenges.  
43 *International Cataloguing and Bibliographic Control (ICBC)*, 36, 78-82.  
44  
45 ZICKUHR, K. & RAINIE, L. 2011. *Wikipedia, past and present* [Online]. Pew Research Center. Available:  
46 <http://www.pewinternet.org/2011/01/13/wikipedia-past-and-present/> [Accessed May 2017].  
47  
48 ZIPF, G. K. 1949. *Human Behaviour and the Principle of Least-Effort*. Cambridge MA edn. Addison-Wesley, Reading.  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# Multiagent systems : algorithmic, game-theoretic, and logical foundations

1  
2 Author: [Yoav Shoham](#); [Kevin Leyton-Brown](#)  
3  
4 Publisher: Cambridge ; New York : Cambridge University Press, 2009.  
5  
6 Edition/Format: Book Computer File :  
7 English [View all editions and formats](#)  
8  
9 Database: WorldCat  
10  
11 Summary: This is an introduction to a burgeoning interdisciplinary field, with an emphasis on foundational material.  
12  
13  
14 Rating: (not yet rated) with reviews - Be the first.  
15  
16 Subjects: [Multiagent systems.](#)  
17 [Electronic data processing – Distributed processing.](#)  
18 [Mehragentensystem.](#)  
19

Library Hi Tech




WIKIPEDIA  
The Free Encyclopedia

Article Talk

## Multi-agent system

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia

Interaction  
Help



WIKIPEDIA  
The Free Encyclopedia

Article Talk

## Distributed computing

From Wikipedia, the free encyclopedia

"Distributed Information Processing" redirects here.

**Distributed computing** is a field of [computer science](#) where [computers](#) communicate and coordinate their actions in order to solve a problem that no single computer could solve. Distributed systems vary from [SOA-based systems](#) to [peer-to-peer systems](#). A [computer program](#) that runs in a distributed system is called a [distributed program](#). Distributed computing also refers to the use of distributed systems to solve a problem that no single computer could solve. <sup>[3]</sup> which.com

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia

Interaction  
Help

20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34



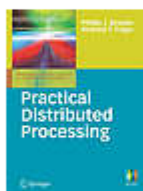
su:Electronic data processing Distributed processing

Advanced Search Find a Library

Search results for 'su:Electronic data processing Distributed processing'

Results 1-10 of about 11,107 (.11 seconds) << First

Select All Clear All Save to: [New List] Save Sort by: Relevance

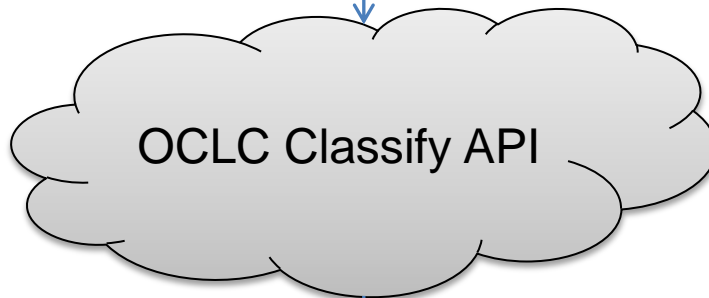
1.  **Practical distributed processing**  
by Phillip J Brooke; Richard F Paige  
eBook : Document [View all formats and languages](#) »  
Language: English  
Publisher: London : Springer, ©2008.  
Database: WorldCat  
[View all editions](#) »

2. **Distributed processing systems**  
by Robert J Thierauf  
Print book : [View all formats and languages](#) »

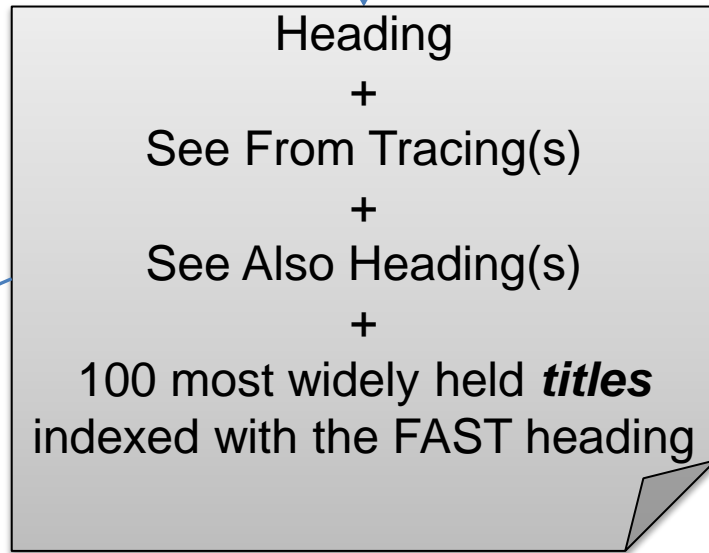
### FAST Records

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16

Control Number:  
Heading:  
See From Tracing(s):  
See Also Heading(s):  
WC Subject Usage:



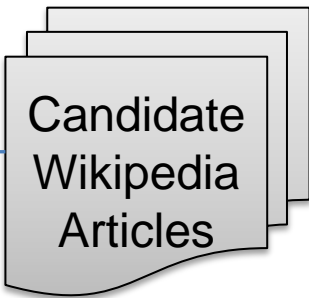
OCLC Classify API



Heading  
+  
See From Tracing(s)  
+  
See Also Heading(s)  
+  
100 most widely held **titles**  
indexed with the FAST heading



WikipediaMiner

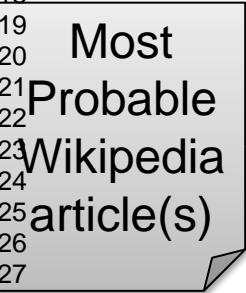


Candidate  
Wikipedia  
Articles



WEKA

17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30



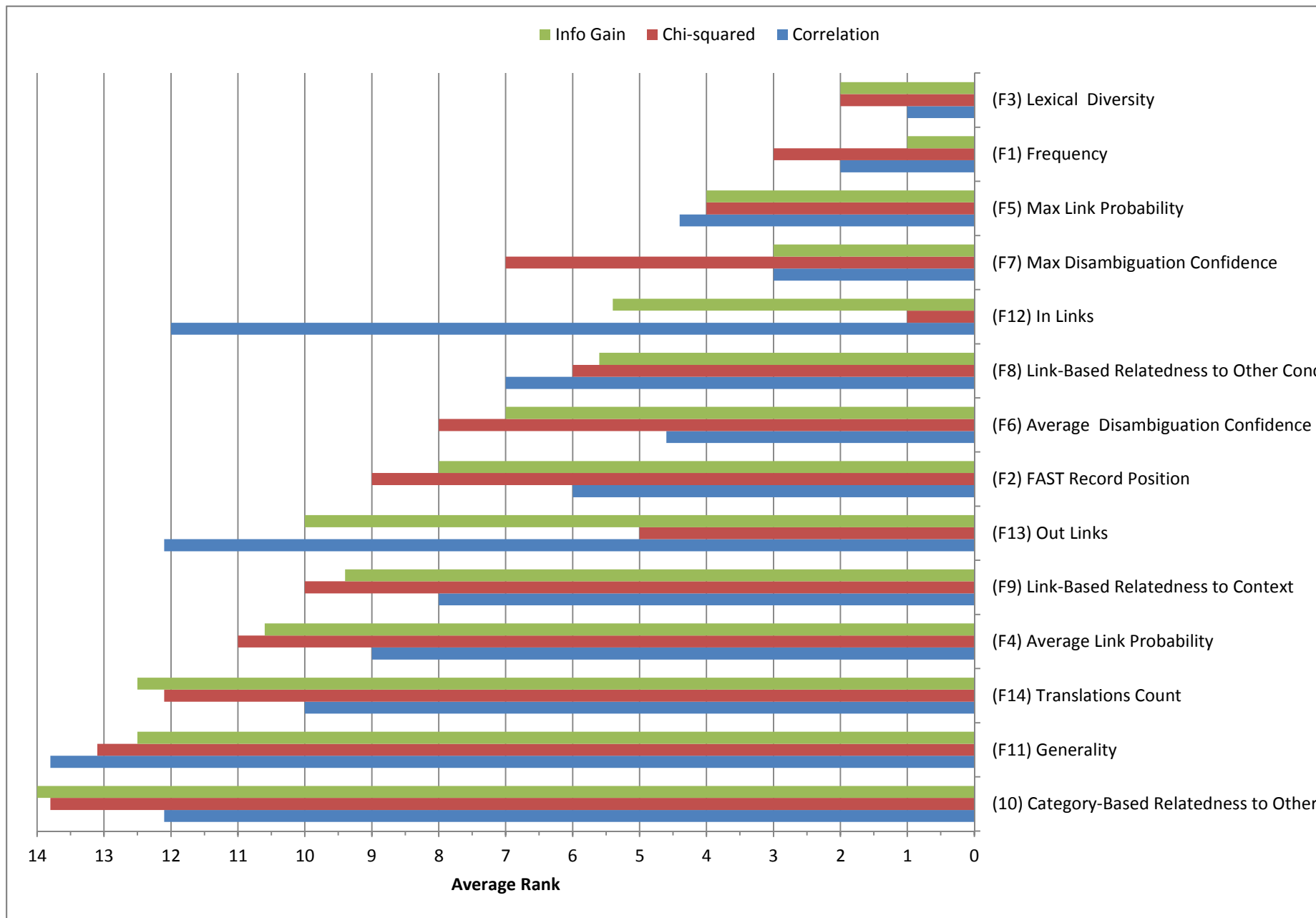
Most  
Probable  
Wikipedia  
article(s)

- Page 15 of 21
1. The invisible cure : Africa, the West, and the fight against **AIDS**
  2. **AIDS education and prevention**
  3. The wisdom of whores : bureaucrats, brothels, and the business of **AIDS**
  4. Love is the cure : on life, loss, and the end of **AIDS**
  5. A young man's guide to sex
  6. CRISIS : **heterosexual behavior** in the age of **AIDS**
  7. What you can do to avoid **AIDS**
  8. The age of **AIDS**
  9. Epidemics : opposing viewpoints
  10. **Confronting AIDS**: directions for public health,health care,and research
  11. **AIDS : sexual behavior and intravenous drug use**
  12. The **spread of AIDS**
  13. Moving mountains : the race to treat global **AIDS**
  14. **Confronting AIDS.**
  15. Rx for survival : why we must rise to the global health challenge
  16. **AIDS** challenge : **prevention education** for young people
  17. **AIDS** : policies and programs for the workplace
  18. Advice for life : a woman's guide to **AIDS risks and prevention**
  19. Understanding and **preventing AIDS**
  20. Women and **AIDS** : negotiating safer practices, care, and representation
  21. **AIDS** and patient management : legal, ethical, and social issues
  22. Behavioral aspects of **AIDS**

23. **Preventing AIDS** : the design of effective programs
24. **Positive prevention : reducing HIV transmission** among people living with **HIV/AIDS**
25. Letting them die : why **HIV/AIDS intervention** programmes fail
26. Primary **prevention of AIDS** : psychological approaches
27. **AIDS,behavior,and culture: understanding evidence-based prevention**
28. Global **AIDS** : myths and facts : tools for fighting the **AIDS pandemic**
29. Integrating cultural, observational, and epidemiological approaches in **the prevention of drug abuse and HIV/AIDS**
30. **AIDS** : a health care management response
31. Evaluation and management of early **HIV infection.**
32. **AIDS, drugs, and prevention** : perspectives on individual and community
33. Rethinking **AIDS prevention** : learning from successes in developing countries
34. **AIDS** : effective health communication for the 90s
35. Innovative approaches to health psychology : **prevention and treatment lessons from AIDS**
36. After the cure : managing **AIDS** and other public health crises
37. How effective is **AIDS education?**.
38. Responding to the **AIDS epidemic**
39. **Preventing AIDS** in drug users and their sexual partners
40. Denying **AIDS** : conspiracy theories,pseudoscience, and human tragedy



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49



**Table 1.** Numeric values for the FAST Record Position.

FAST heading (preferred label)	See from tracings (alternative labels)	See also headings (related terms)	FAST Record Position (numeric value)
✓	✓	✓	7
✓	✓	X	6
✓	X	✓	5
✓	X	X	4
X	✓	✓	3
X	✓	X	2
X	X	✓	1
X	X	X	0

Library Hi Tech

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Table 2. Sample FAST to Wikipedia mappings.

FAST Heading	Wikipedia Article(s)	WorldCat Usage
APL (Computer program language)	APL (programming language)	878
Aboriginal Australian literature	Indigenous Australians	53
	Aboriginal Australians	
	Australian literature	
Accordion and percussion music	Accordion	34
	Percussion instrument	
Abortion -- Complications	Unsafe abortion	149
	Abortion	
AN/BSY-2 (Computer system)	n/a	2
Abortion -- Moral and ethical aspects	Abortion debate	2687
	Religion and abortion	
Abbadides	Abbadid dynasty	14
Abaza	Abazins	23
Aboriginal Tasmanians -- Mixed descent	Aboriginal Tasmanians	5
Abdominal aorta -- Radiography	Abdominal aorta	8
	Aortography	
Abdomen -- Tumors	Abdominal cavity	115
	Neoplasm	
Abdomen -- Wounds and injuries	Abdominal trauma	149
Acaricides -- Physiological effect	Acaricide	8
AIA Gold Medal	AIA Gold Medal	7
Abdominal aorta -- Surgery	Abdominal aorta	21

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Table 3.** Classification performance achieved using various classification algorithms in Weka.

Classifier (Weka implementation)	Category	Precision	Recall	F <sub>1</sub>	Seconds taken to build model
Logistic Regression (logistic)	Corresponding	0.752	0.502	0.602	1.59
	Non-Corresponding	0.997	0.999	0.998	
	Weighted Average	0.996	0.996	0.996	
Multilayer Perceptron (MultilayerPerceptron)	Corresponding	0.735	0.577	<b>0.647</b>	77.95
	Non-Corresponding	0.998	0.999	<b>0.998</b>	
	Weighted Average	0.996	0.997	<b>0.996</b>	
Decision Tree (J48)	Corresponding	0.738	0.515	0.606	1.69
	Non-Corresponding	0.997	0.999	0.998	
	Weighted Average	0.996	0.996	0.996	
Random Forest (RandomForest)	Corresponding	0.844	0.473	0.606	23.26
	Non-Corresponding	0.997	1.000	0.998	
	Weighted Average	0.996	0.997	0.996	
Multilayer Perceptron + Feature Selection all features except F10	Corresponding	0.696	0.560	0.621	69.32
	Non-Corresponding	0.998	0.999	0.998	
	Weighted Average	0.996	0.996	0.996	

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Table 4.** “FAST Record Position” feature values of the “corresponding” concepts in the dataset.

FAST heading (preferred label)	See from tracings (alternative labels)	See also headings (related terms)	FAST Record Position (numeric value)	Total number of instances in the “corresponding” category (%)
✓	✓	✓	7	0 (0%)
✓	✓	X	6	101 (42%)
✓	X	✓	5	0 (0%)
✓	X	X	4	116 (48%)
X	✓	✓	3	0 (0%)
X	✓	X	2	4 (1.7%)
X	X	✓	1	0 (0%)
X	X	X	0	20 (8.3%)

Library Hi Tech

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Table 5.** Number of Wikipedia articles citing valid ISBNs.

<b>1 or more</b>	<b>1-3</b>	<b>3-6</b>	<b>6-12</b>	<b>More than 12</b>
375,138	321,293	32,007	15,064	6,774

Library Hi Tech