Fostering Library-Wikipedia Integration: Automatic Mapping of FAST Subject Headings to Wikipedia Articles

Abdulhussain E. Mahdi & Arash Joorabchi

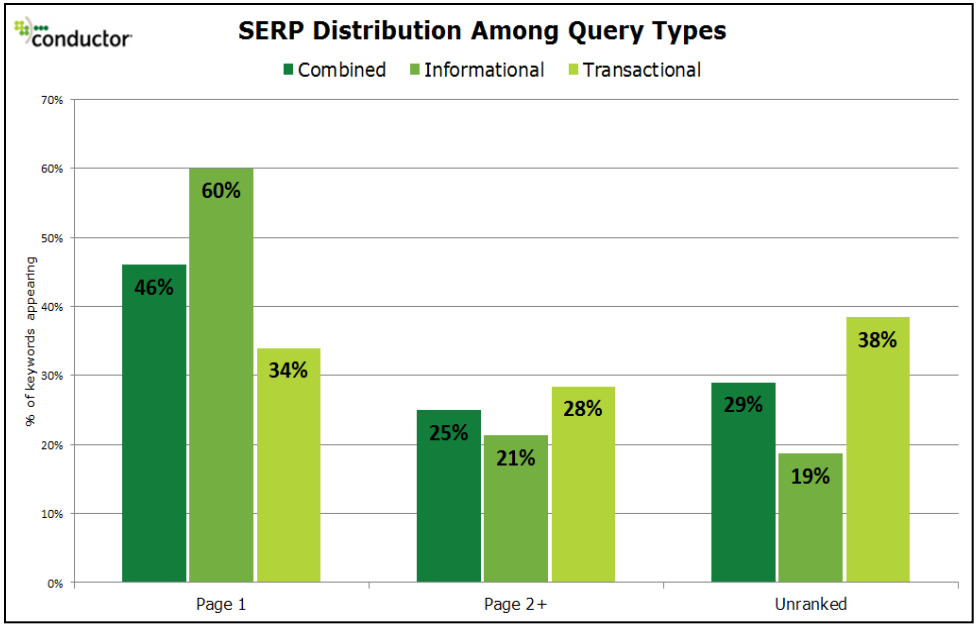
Department of Electronic and Computer Engineering
University of Limerick, Ireland

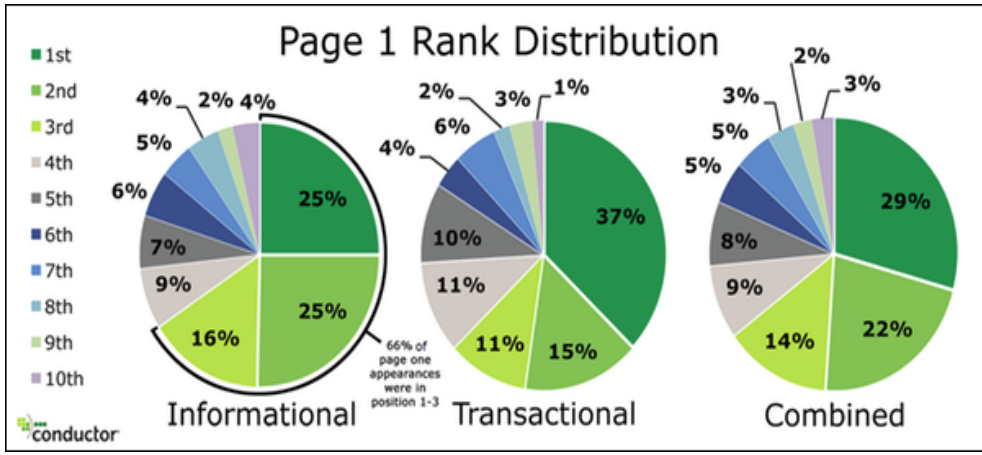
Supported by: OCLC® alise

UNIVERSITY of LIMERICK
OLLSCOIL LUIMNIGH

Google-Wikipedia Information Seeking Paradigm

- Few information searches (~1%) start from library websites
- The great majority of information seeking activities (84%) start from search engines such as Google.
- Google-Wikipedia is becoming a prevalent information seeking paradigm in which the information seeker submits an informational query to Google and then follows one of the search results redirecting to a relevant article on Wikipedia.
- 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.

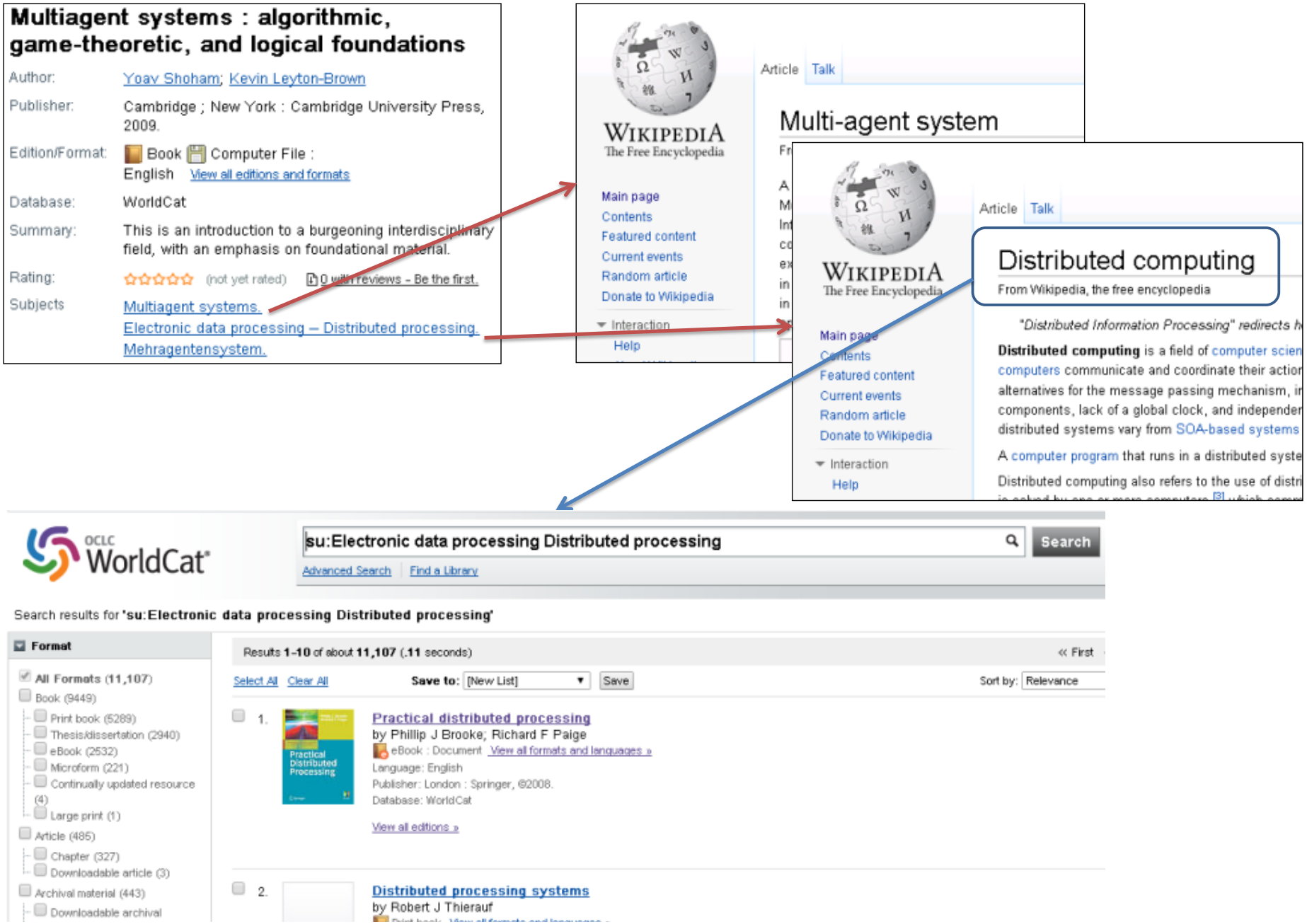




Pictures courtesy of *http://www.conductor.com/blog/2012/03/wikipedia-in-the-serps-appears-on-page-1-for-60-of-informational-34-transactional-queries/*

A Vision for Wikipedia-Library Integration

- Integrating two major silos of knowledge.
- Creating a bi-directional link and flow of information and users between the Wikipedia and libraries.
- Information seekers may start their search activities from either of these sources and traverse back and forth as needed.
- Users browsing a library catalogue would be able to see the subject metadata of each item in form of a set of FAST subject headings which are linked to their equivalent Wikipedia articles.
- A user who reaches a Wikipedia article of a topic via conducting a Google or Wikipedia search, will be provided with a link to the article's equivalent FAST subject heading(s) on the *WorldCat.org* website.


Multiagent systems : algorithmic, game-theoretic, and logical foundations
Author: Yoav Shoham; Kevin Leyton-Brown
Publisher: Cambridge ; New York : Cambridge University Press, 2009.
Edition/Format: Book Computer File : English View all editions and formats
Database: WorldCat
Summary: This is an introduction to a burgeoning interdisciplinary field, with an emphasis on foundational material.
Rating: (not yet rated) 0 with reviews - Be the first.
Subjects
Multiagent systems.
Electronic data processing -- Distributed processing.
Mehragentensystem.
WIKIPEDIA
The Free Encyclopedia
Article Talk
Multi-agent system
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Interaction
Help
Distributed computing
From Wikipedia, the free encyclopedia
Distributed computing is a field of computer scien
computers communicate and coordinate their actio
alternatives for the message passing mechanism, i
components, lack of a global clock, and independer
distributed systems vary from SOA-based systems
A computer program that runs in a distributed syste
Distributed computing also refers to the use of distri
OCLC WorldCat
su:Electronic data processing Distributed processing
Search
Advanced Search Find a Library
Search results for 'su:Electronic data processing Distributed processing'
Format
All Formats (11,107)
Book (9449)
Print book (5289)
Thesis/dissertation (2940)
eBook (2532)
Microform (221)
Continually updated resource (4)
Large print (1)
Article (485)
Chapter (327)
Downloadable article (3)
Archival material (443)
Downloadable archival
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

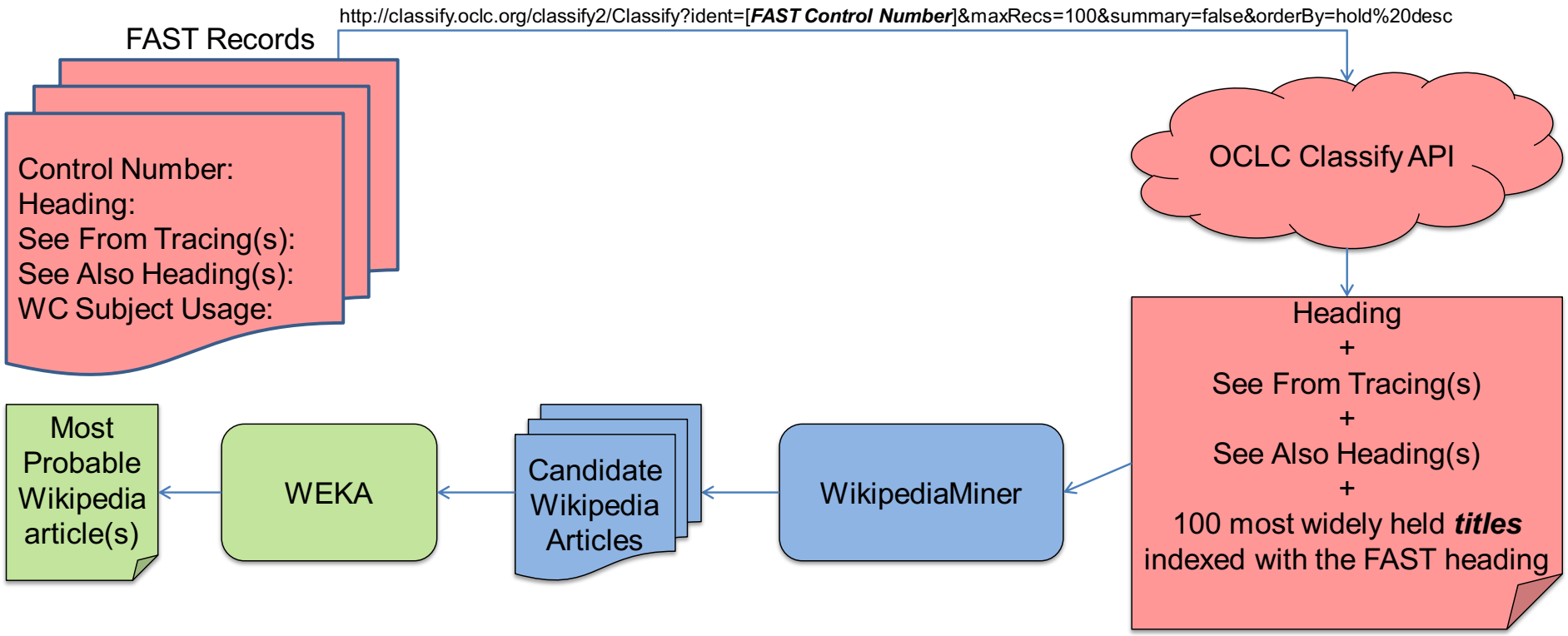
Automatic Mapping of FAST Subject Headings to Wikipedia Articles

- 1.7 million FAST subject headings are divided into 8 different facets (personal names, corporate names, geographic names, events, titles, time periods, topics, and form/genre).
- The initial focus of our project is on mapping the 400,000 topical subject headings (MARC Field 650) to their corresponding Wikipedia articles.
- Establishing such mapping would be the most fruitful in terms of realising the proposed vision of full library-Wikipedia integration.
- The English version of Wikipedia currently contains over 5 million articles which their equivalent FAST headings could belong to any of the 8 facets of FAST.

LCSH/FAST (MARC-650)

WIKIPEDIA
The Free Encyclopedia

Automatic Mapping Method



1. **Data collection:** retrieving titles of library materials indexed with the given FAST heading.
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".

Stage 1: Data Collection

- Retrieve titles of the books indexed with the FAST heading to be mapped, using the OCLC Classify API:

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

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

Top 40 titles indexed with the FAST heading "AIDS (Disease)—Prevention"

Stage 2: Candidate Detection

- Detecting all the candidate Wikipedia articles/concepts appearing in the "titles" file of the FAST headings, using an open-source toolkit called Wikipedia-Miner.
- E.g., 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, ...

- A significant number of detected Wikipedia concepts would be related to the FAST heading, however, only one or a few directly correspond to the given FAST heading:
  - **One-to-one mapping:** "**AIDS (Disease)–Prevention**" ➔ **Prevention of HIV/AIDS**
  - **One-to-many mapping:** "**Abnormalities, Human--Genetic aspects**" **has two corresponding articles in Wikipedia: "Congenital disorder" and "Genetic disorder".**

Stage 3: Candidate Classification

- This stage involves finding the most probable corresponding Wikipedia article(s) for the FAST heading among the large set of candidates detected in the FAST heading's "titles" file.
- Using a Machine Learning (ML) based binary classifier each candidate concept is classifies as either "corresponding" or "non-corresponding".

- To build and train such an ML-based 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.

  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.a) Features for Candidate Wikipedia Concepts

- We have devised a set of 14 positional, statistical, and semantic features to capture and reflect various characteristics of those candidates which have the highest probability of belonging to the "corresponding" category

1. Frequency

2. FAST Record Position

3. Lexical Diversity

4. Average Link Probability
5. Max Link Probability

6. Average Disambiguation Confidence
7. Max Disambiguation Confidence

8. Link-Based Relatedness to Other Concepts
9. Link-Based Relatedness to Context
10. Category-Based Relatedness to Other Concepts

11. Generality

12. In Links
13. Out Links

14. Translations Count

3.b) Building a Training & Testing Dataset

- A dataset of manually mapped FAST Headings-to-Wikipedia Concepts instances, is fed to an ML-based classification algorithm for learning a prediction model from.
- The dataset was built by manually mapping a set of 200 randomly chosen FAST headings to their equivalent Wikipedia concepts/articles.
- 110 of the total 200 sample headings were mapped to their single corresponding Wikipedia articles (i.e., one-to-one mappings)
- 60 were mapped to multiple articles (i.e., one-to-many mappings),
- For the remaining 30 headings examined, no corresponding Wikipedia articles were found (i.e., out-of-date and/or very specific concepts such as: “AN/BSY-2 (Computer system), WorldCat usage: 2”
- The final dataset contains a total of 170 FAST headings manually mapped to 241 Wikipedia articles.
- Each FAST heading in the dataset is assigned an average of 267 candidate Wikipedia articles out of which only 1.1 belong to the “corresponding” category and the rest belong to the “non-corresponding”.

Sample Mappings From the Dataset

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

Experimental Results & Evaluation (1)

- Evaluation results of our experiments with various well-known ML-based classification algorithms, measured using standard information retrieval metrics and 10-fold cross-validation.

Table 1. Classification performance achieved using various classification algorithms in Weka.

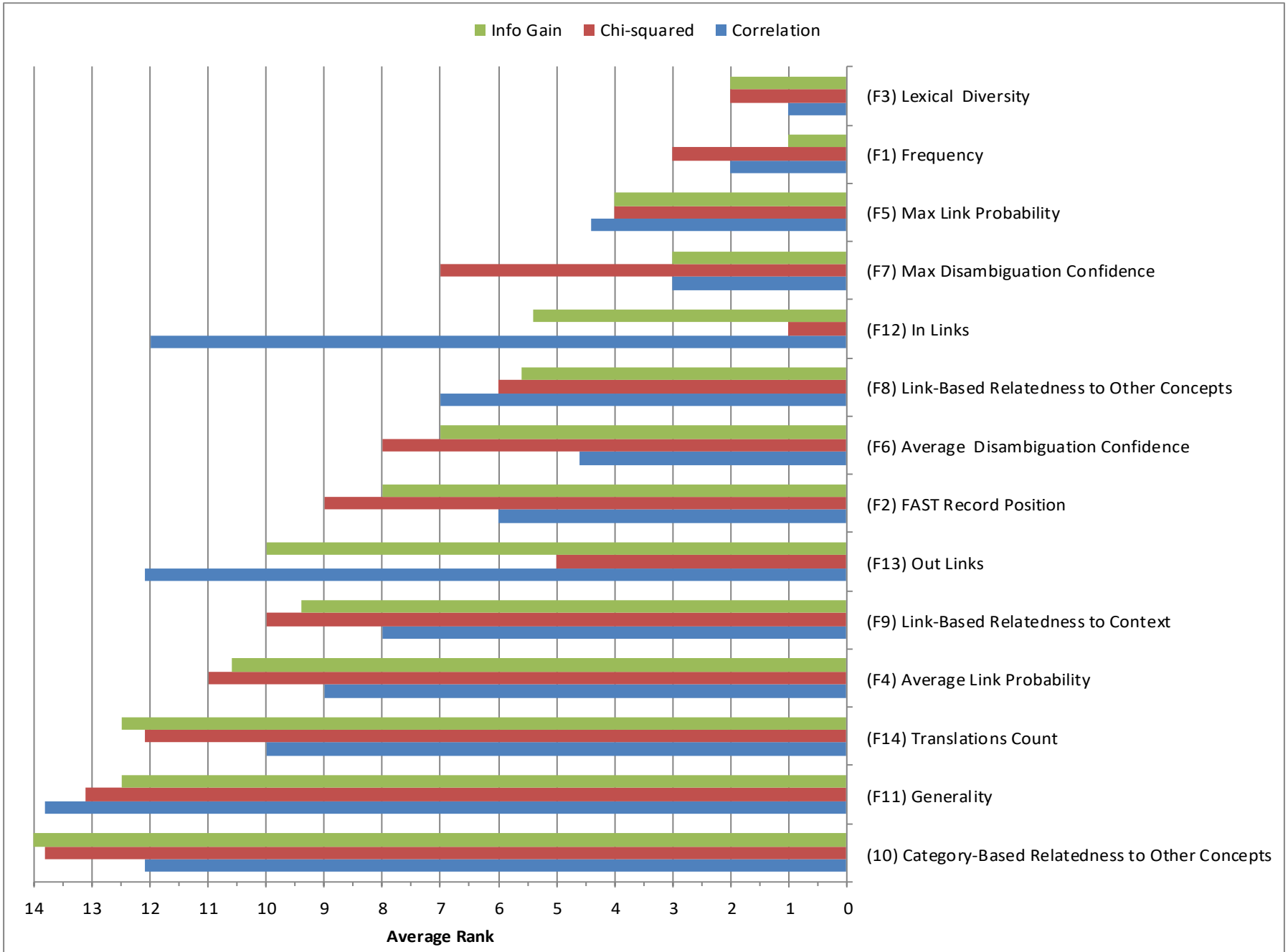
Classifier (Weka implementation)	Category	Precision	Recall	F_1
Logistic Regression (logistic)	Corresponding	0.752	0.502	0.602
	Non-Corresponding	0.997	0.999	0.998
	Weighted Average	0.996	0.996	0.996
Multilayer Perceptron (MultilayerPerceptron)	Corresponding	0.735	0.577	**0.647**
	Non-Corresponding	0.998	0.999	**0.998**
	Weighted Average	0.996	0.997	**0.996**
Decision Tree (J48)	Corresponding	0.738	0.515	0.606
	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
	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
	Non-Corresponding	0.998	0.999	0.998
	Weighted Average	0.996	0.996	0.996

$$Pr = \frac{\text{Number of correctly mapped FAST terms}}{\text{Total mapped}} = \frac{TP}{TP + FP}$$

$$Re = \frac{\text{Number of correctly mapped FAST terms}}{\text{Total possible correct}} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Pr \times Re}{Pr + Re}$$

Experimental Results & Evaluation (2)



Future Work

- Optimizing the Wikipedia-Miner to reduce the number of candidate concepts per FAST heading.
- Analysing the textual content of 4,799,116 Wikipedia articles from the English Wikipedia used in this study, showed that a considerable number of articles (375,138) contain at least one valid ISBN number in their "References" section.

Table 2. Number of Wikipedia articles citing valid ISBNs.

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

- Using a citation analysis-based technique we could utilise these currently existing links between Wikipedia articles and library resources to enhance our proposed mapping method and potentially improve its accuracy.
- Demonstrate the application of the proposed mapping method by developing a browser plugin capable of redirecting users, when appropriate, from Wikipedia articles to *WorldCat.org* website for further reading on their subject of interest.

Thank You!

Questions…

For more information, please contact:

Hussain.Mahdi@ul.ie Arash.Joorabchi@ul.ie

This work is supported by:

The OCLC/ALISE Library & Information Science Research Grant Program (LISRGP) 2016

OCLC® alise UNIVERSITY of LIMERICK OLLSCOIL LUIMNIGH