

A k-Nearest-Neighbour Method for
Classifying Web Search Results
with Data in Folksonomies

by Ching-man Au Yeung, Nicholas Gibbins, Nigel Shadbolt

Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton

The Problem of Ambiguity

- Queries by ambiguous terms return many irrelevant results
- Example: *bridge*
 - 1) a kind of card games;
 - 2) a form of architectural structure;
 - 3) a design pattern in software development;
 - 4) a device in computer networking

Introduction

The screenshot shows the Delicious website interface. At the top, it says "delicious social bookmarking" and "It's Free! Join Now Sign In". Below this is a blue banner with the text "Your bookmarks will organize themselves. Tag your bookmarks. Collections will naturally emerge." and a "Learn More" link. A search bar is present with the text "Search the biggest collection of bookmarks in the universe...". Below the search bar, there are sections for "Popular Bookmarks" and "Explore Tags". The "Popular Bookmarks" section lists several items with their titles, tags, and the number of bookmarks. For example, "Where to Buy Used Canon Lenses" has 160 bookmarks and tags like "photography", "canon", "used", "shopping", "lenses". Other items include "29 Great Free Textures | Abduzeedo - design inspiration & tutorials" (184 bookmarks), "JeffBridges.com - Ironman book" (149 bookmarks), "Essential free apps for your web design toolkit | News | TechRadar UK" (214 bookmarks), "Toxel.com » 24 Beautiful and Creative Website Headers" (228 bookmarks), "Seadragon Ajax - Microsoft Live Labs" (192 bookmarks), "MAKE: Blog: Arduino Gift Guide!" (160 bookmarks), "27 Free Must-have Online Collaboration Tools - Crazeegeeckchick.com" (328 bookmarks), and "is it going to rain?" (395 bookmarks). A "Popular Tags" section on the right lists various tags like "design", "blog", "videos", "software", "tools", "music", "programming", "webdesign", "reference", "tutorial", "art", "web", "howto", "javascript", "free", "linux", "web2.0", "development", "google", "inspiration", "photography", "news", "food", "flash", "css", "blogs", "education", "business", "technology", and "travel".

Delicious

The screenshot shows the BibSonomy website interface. At the top, it says "BibSonomy :: search:all ::" and "A free social bookmark and publication sharing system." Below this is a navigation bar with links for "tags", "relations", "groups", "popular", "username:", and "password:". There is also a "help · blog · about" and "login · register" section. The main content area is divided into two columns: "bookmarks" and "publications". The "bookmarks" column lists several items with their titles and tags. For example, "Définitions dans le cadre des soins palliatifs" has tags like "design", "blog", "videos", "software", "tools", "music", "programming", "webdesign", "reference", "tutorial", "art", "web", "howto", "javascript", "free", "linux", "web2.0", "development", "google", "inspiration", "photography", "news", "food", "flash", "css", "blogs", "education", "business", "technology", and "travel". The "publications" column lists several items with their titles and tags. For example, "Foundations of Therapeutic Interviewing" has tags like "edge", "economics", "editor", "folksonomy", "free", "history", "howto", "information", "internet", "java", "language", "linux", "maps", "marketing", "math", "mathematics", "metadata", "model", "music", "network", "networks", "netways", "nlp", "of", "ontology", "opensource", "optimization", "php", "programming", "rdf", "repository", "research", "retrieval", "science", "search", "security", "semantic", "semanticweb", "service", "social", "evaluation".

BibSonomy

Collaborative Tagging Systems

- ◆ Aggregate user-contributed metadata of Web resources
- ◆ Provide rich information about the relations between different tags
- ◆ Sources for understanding how keywords are used on the Web

Multiple Meanings of Tags

- ◆ Tags have multiple meanings, or they are used in different contexts
- ◆ It is possible to extract related tags in different contexts
- ◆ E.g. sf:

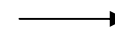
(california, bayarea, travel, ...) → San Francisco
(scifi, fantasy, fiction, ...) → Science Fiction

Our Proposal

- ◆ Building classifiers using data in folksonomies:

Wikipedia page of San Francisco

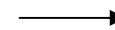
The City and County of San Francisco is the fourth most populous city in **California** and the 14th most populous city in the United States ... Among the most densely populated cities in the country, San Francisco is part of the San Francisco **Bay Area** ... The city is located at the tip of the San Francisco Peninsula, with the Pacific Ocean to the west, ...



San
Francisco

Wikipedia page of Science Fiction

Science fiction (abbreviated SF or **sci-fi** with varying punctuation and capitalization) is a broad genre of **fiction** that often involves speculations based on current or future science or technology. Science fiction is found in books, art, television, films, games, theatre, and other media ... this includes **fantasy**, horror, and related genres.

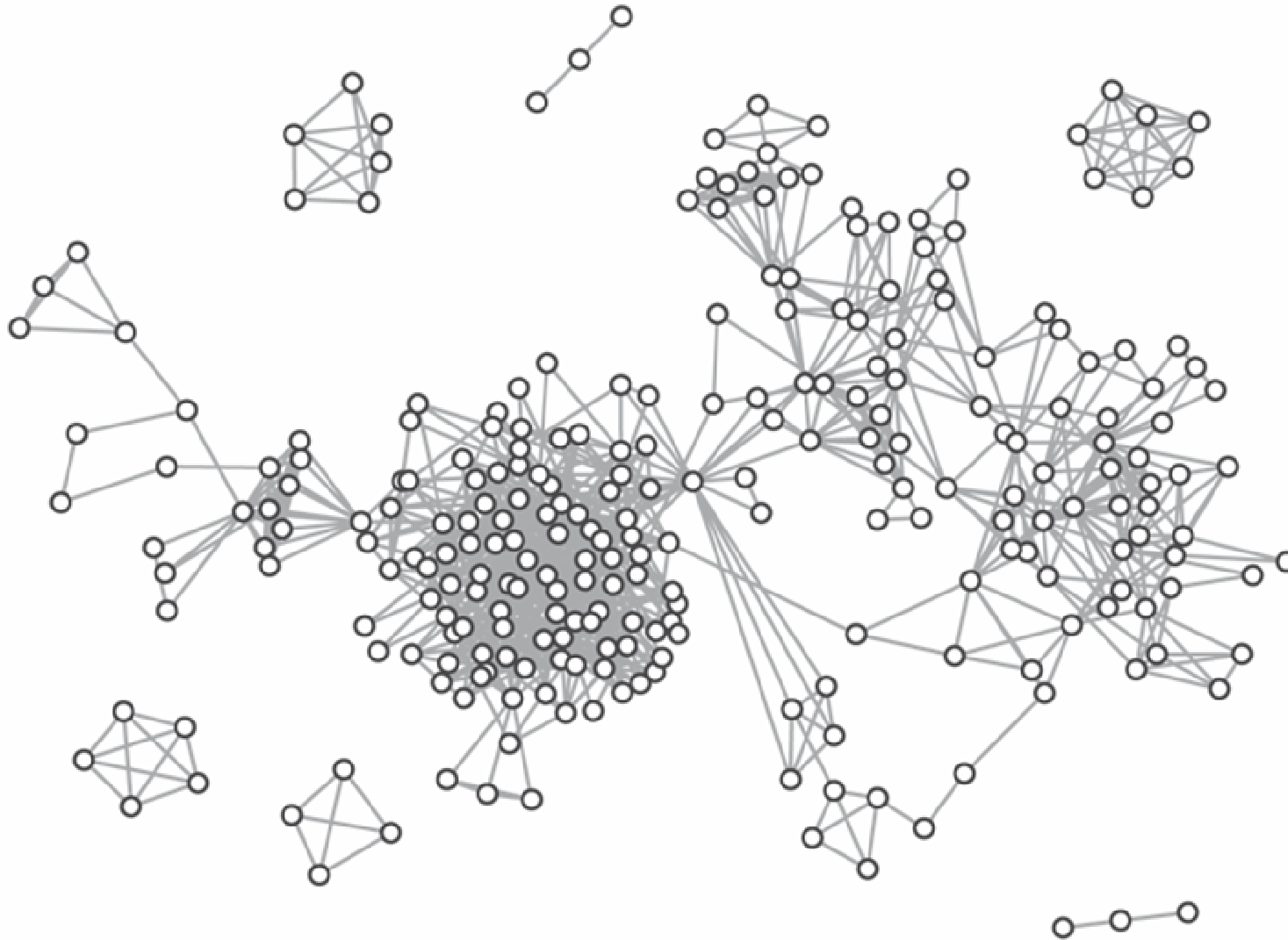


Science
Fiction

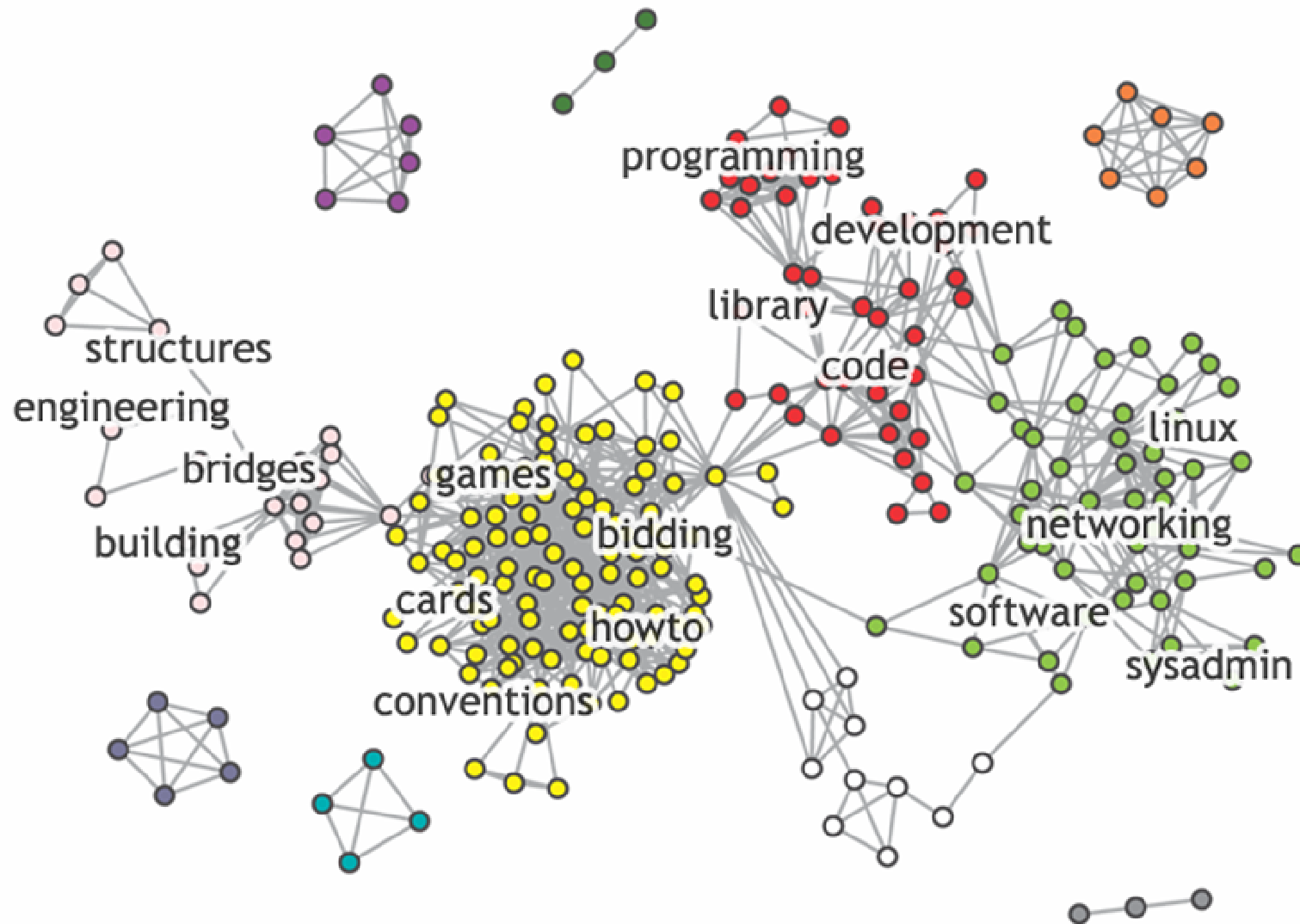
Clustering of Folksonomy Networks

- ◆ Construct a document network from a folksonomy
- ◆ Cluster documents based on the users who have used the tag on the documents
(the community-discovery algorithm described in [Newman 2004] is used in this paper)
- ◆ Extract frequently co-occurred tags as representations of the different classes (meanings)

Clustering of Folksonomy Networks



Clustering of Folksonomy Network



Clustering of Folksonomy Network

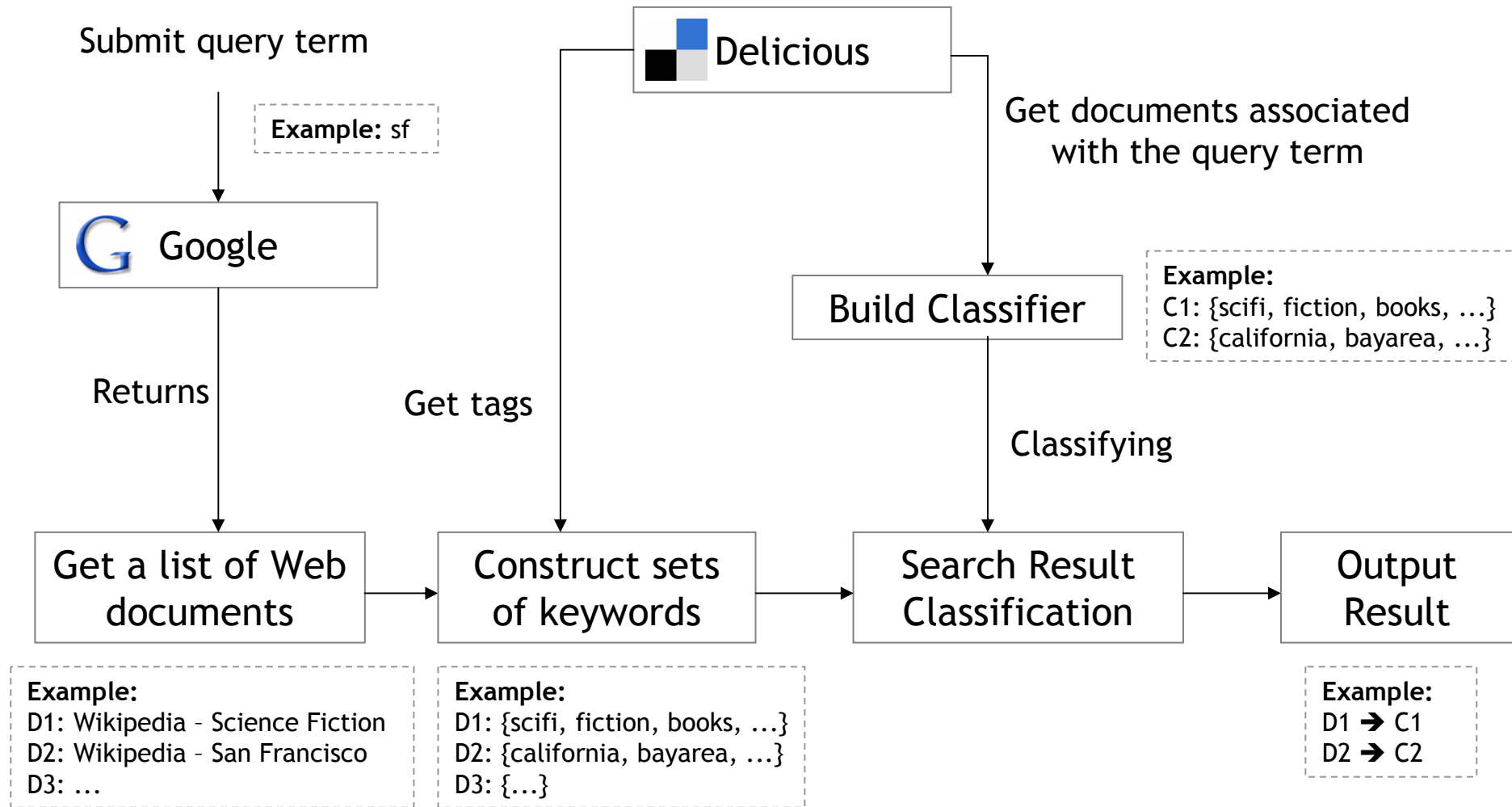
Design pattern bridge, programming, development, library, code, ruby, tools, software, adobe, dev

Card game bridge, games, cards, game, imported, howto, conventions, card, bidding, online

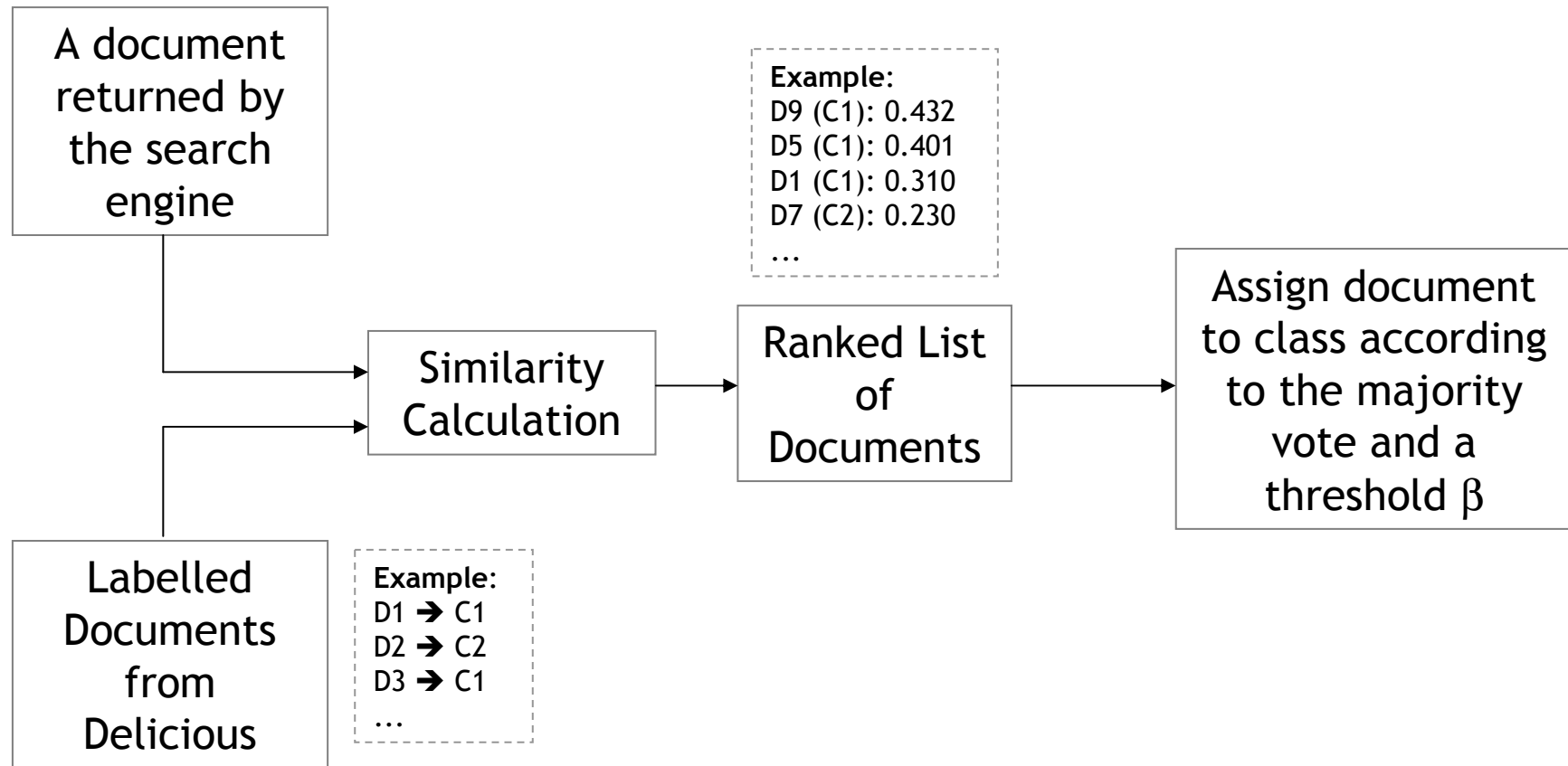
Computer networking bridge, networking, linux, network, howto, software, sysadmin, firewall, virtualization, security

Architecture bridge, bridges, structures, engineering, science, physics, school, education, building, reference

Web Search Result Classification



k-Nearest-Neighbour Classifier



Data Preparation

- ◆ Ten tags which are used in multiple contexts in Delicious are chosen
- ◆ Documents associated with the tags as well as the users who assigned the tags are retrieved
- ◆ Testing dataset obtained by submitting query to Google and obtaining the top 50 pages returned

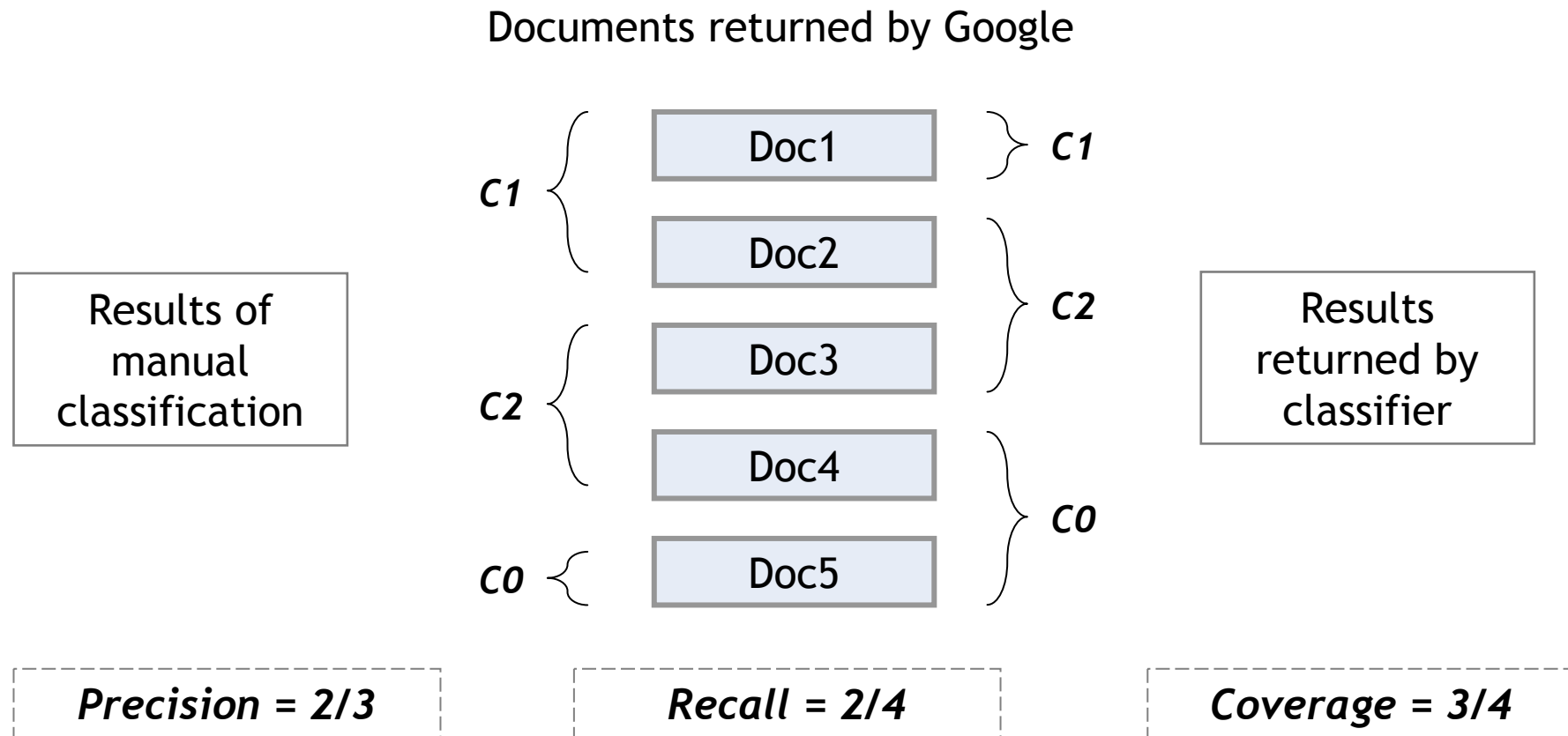
Classifiers Built

Tag	Context	Class Label
sf	San Francisco	sanfrancisco, bayarea, san, francisco, california, travel, events
	Science fiction	scifi, fiction, books, sci-fi, literature, writing, science, fantasy
soap	Cleaning agent	soapmaking, diy, recipes, crafts, shopping, making, howto
	Web services	webservices, webservice, programming, web, xml, soa, java
wine	Software application	linux, ubuntu, howto, windows, software, tutorial, emulation
	Beverage	food, shopping, drink, vino, cooking, alcohol, blog, news
xp	Windows XP	windows, software, tools, pc, computer, tech, winxp, microsoft
	Extreme programming	software, programming, process, methodology, development

Performance Measures

- ◆ ***Precision***
Measures the percentage of documents which are classified correctly.
- ◆ ***Recall***
Measures the percentage of classifiable documents which are classified correctly
- ◆ ***Coverage***
Measures the proportion of documents the classifiers are able to classify

An Example



Evaluation

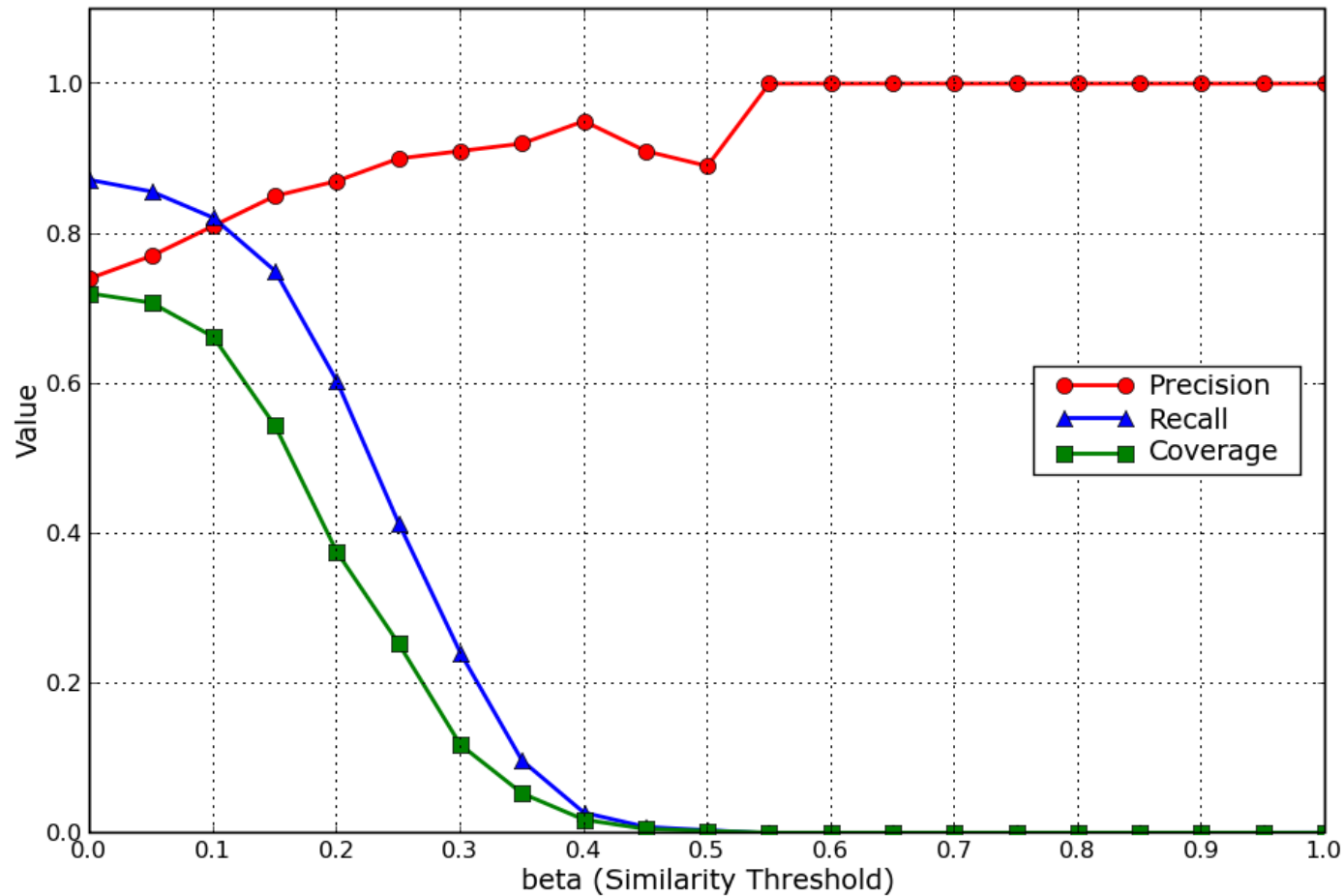
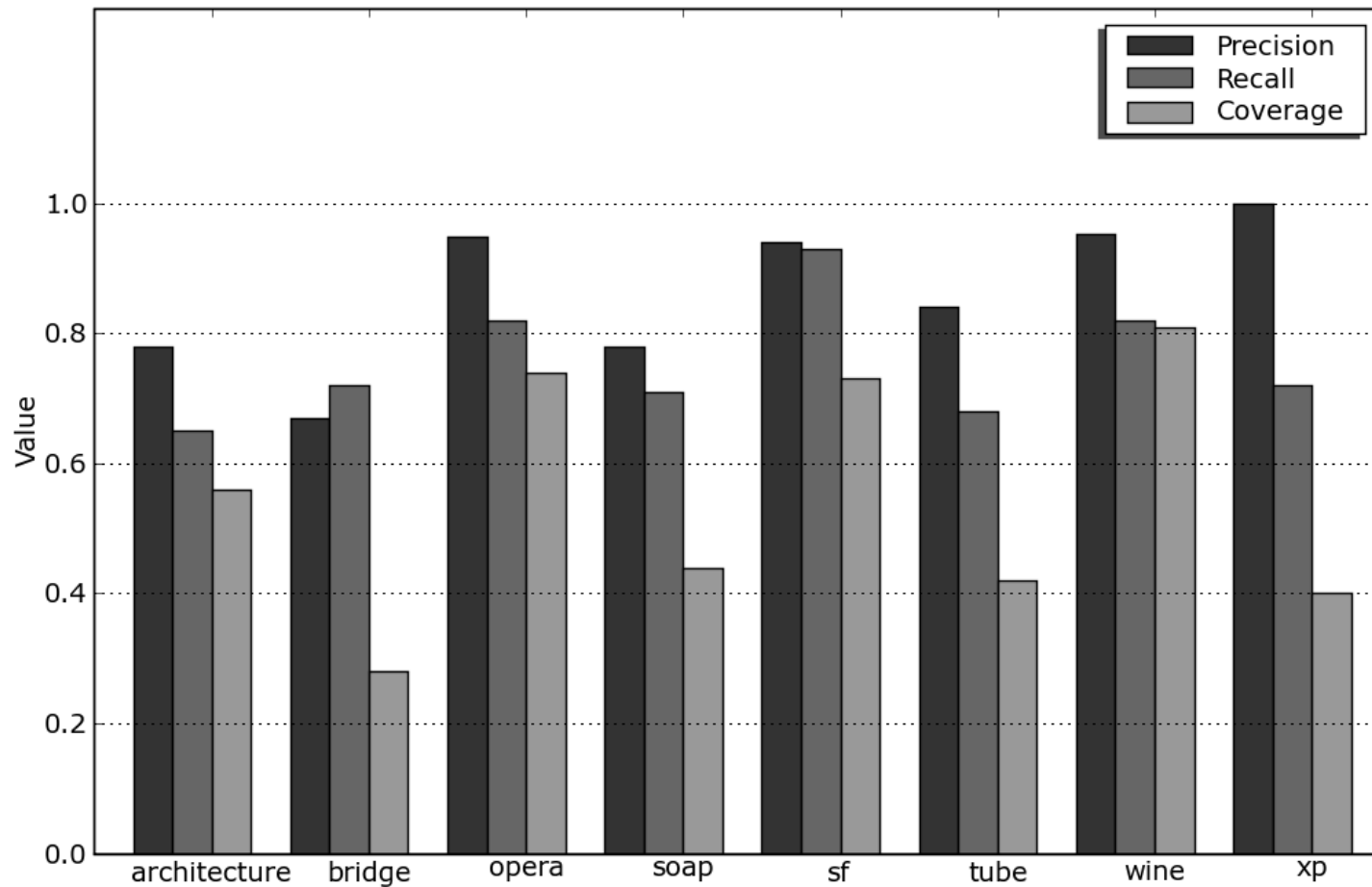


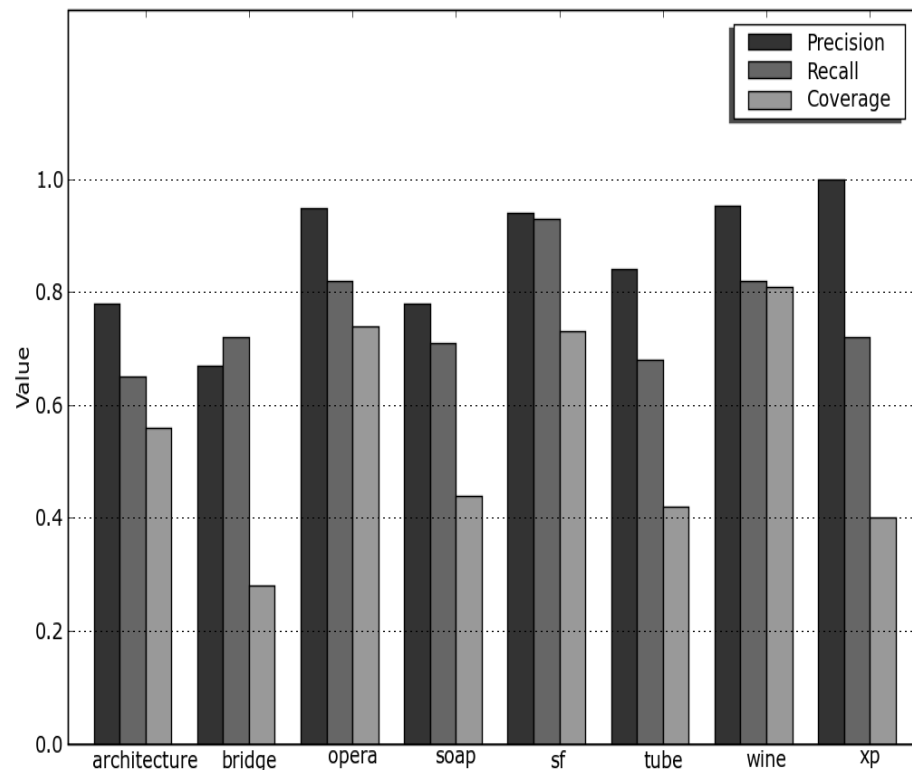
Figure 1. Precision, recall and coverage against different values of β .

Evaluation



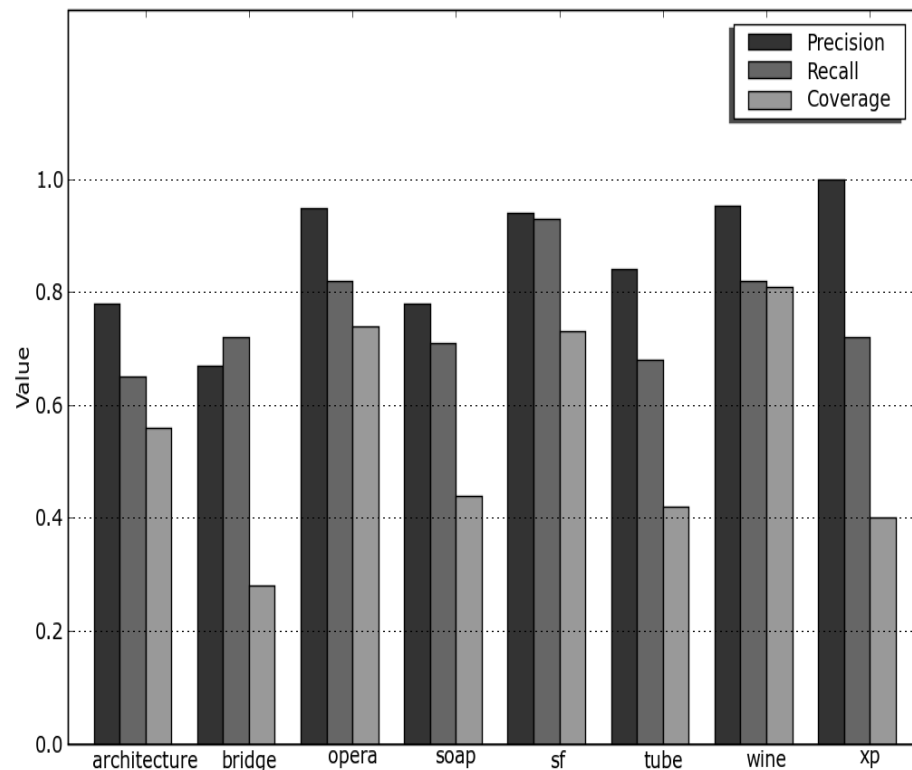
**Figure 2. Precision, recall and coverage for different tags.
($k = 11$, $\beta = 0.15$)**

Precision



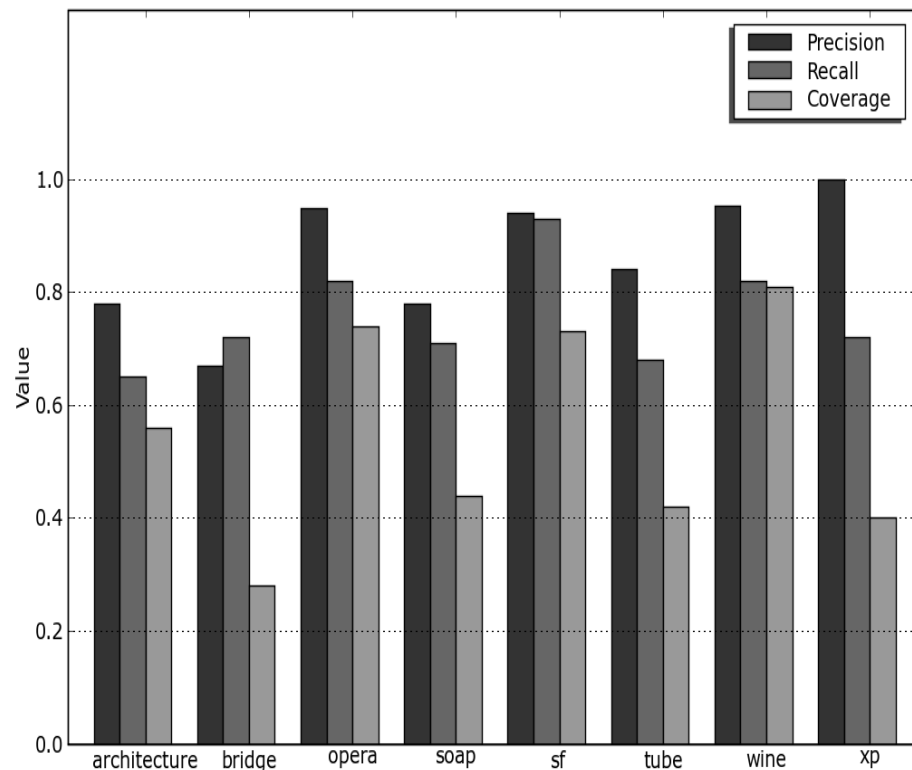
- ◆ Precision is generally quite high (67-100%)
- ◆ Clustering process provides good basis for the kNN classifier
- ◆ Low precision cases: different keywords even for same context

Recall



- ◆ Recall ranges from 65% to 93%
- ◆ Some documents cannot be classified (recognised)
- ◆ Mainly because that keywords do not match well

Coverage



- ◆ Has the largest range: 28-81%
- ◆ Due partly to low recall
- ◆ Some contexts not discovered by the clustering process (e.g. tube)
- ◆ Also, there are irrelevant results (e.g. bridge)

C onclusions

- ◆ Folksonomies offer rich information on the relations and semantics of tags, and can be used to enhance Web search
- ◆ Advantages over using of dictionaries or thesauruses (able to keep up with new meanings)
- ◆ Future research directions:
 1. Building more comprehensive classifiers
 2. Use of other clustering methods
 3. Larger scale of evaluation

Thank You!

Albert Au Yeung
cmay06r@ecs.soton.ac.uk
<http://users.ecs.soton.ac.uk/cmay06r/>