

# Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources

Gordon W. Paynter  
The INFOMINE Project  
Science Library, University of California  
Riverside, CA 92517-5900  
+1 951 827 2279

paynter@library.ucr.edu

## ABSTRACT

This paper describes the development of practical automatic metadata assignment tools to support automatic record creation for virtual libraries, metadata repositories and digital libraries, with particular reference to library-standard metadata. The development process is incremental in nature, and depends upon an automatic metadata evaluation tool to objectively measure its progress. The evaluation tool is based on and informed by the metadata created and maintained by librarian experts at the INFOMINE Project, and uses different metrics to evaluate different metadata fields. In this paper, we describe the form and function of common metadata fields, and identify appropriate performance measures for these fields. The automatic metadata assignment tools in the iVia virtual library software are described, and their performance is measured. Finally, we discuss the limitations of automatic metadata evaluation, and cases where we choose to ignore its evidence in favor of human judgment.

## Categories and Subject Descriptors

*H.3.6 [Information Storage and Retrieval]:* Library Automation

**General Terms:** Algorithms, Measurement, Performance.

## Keywords

Metadata, automatic metadata assignment, automatic metadata evaluation, INFOMINE, iVia.

## 1. INTRODUCTION

Automatic metadata assignment is a challenging research area where academic techniques such as classification and document understanding are applied in practical tools that meet the needs of information seekers by describing and cataloging previously-unexplored resources. Many promising new classifiers have been developed specifically to meet these needs, but all too often their elegance and effectiveness gets lost in the complexities and idiosyncrasies of everyday metadata, and we face the choice of persevering with novel algorithms, or of abandoning them in favor of old friends like Naive Bayes.

At the same time, automatic metadata assignment tools can be simple, almost trivial, and the challenge to the user or developer is to make seemingly-unimportant decisions about humble assignment schemes that are guided by standing rule sets. Should this Keyword be blacklisted? Which Title field best summarizes

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*JCDL '05*, June 7–11, 2005, Denver, Colorado, USA

Copyright 2005 ACM 1-58113-876-8/05/0006...\$5.00.

this document? Should we assign this Subject Heading when the classifier is only 85% sure it is correct?

In our experience, the users and developers of metadata assignment tools make choices like these—both big and small—on a daily basis. Design decisions of this type are made either by the user who encounters them, the developer who is assigned to resolve them, or a subject expert whose judgment is valued. They tend to be made on the basis of few examples, and though some thought is given to the effects of the decision, the wider consequences cannot always be anticipated.

The INFOMINE Project automatically generates metadata for thousands of Internet resources every day, using metadata assignment tools from the iVia Virtual Library Software [27]. The tools range in complexity from simple rules for assigning Title and Creator metadata by harvesting the text of HTML tags, to Keyphrase and Description extraction algorithms based on syntactic processing, to complex Library of Congress Classification (LCC) and Library of Congress Subject Heading (LCSH) classifiers based on machine learning algorithms like Support Vector Machines and Logistic Regression.

This paper describes how INFOMINE developed a metadata evaluation tool in parallel with its metadata assignment tools to support an iterative development process. The evaluations are convenient enough to run at any time, and flexible enough to measure the effectiveness of each assignment process using metrics tailored to the type of metadata. They illustrate the performance of each assignment tool. Most tools exhibit gradual improvement over time, punctuated by sudden improvements when a specific tool is reimplemented, and occasional steps backward when the recommendations of the automatic processes are overruled by human experts.

The paper is arranged as follows. The next section provides background on INFOMINE, and on metadata assignment and evaluation. We then explain our automatic metadata evaluation program, and the statistics it uses. The next section describes each of the metadata assignment tools used in iVia, and assesses their performance with the automatic evaluation tool. We conclude by discussing limitations and future work.

## 2. BACKGROUND

INFOMINE<sup>1</sup> is a virtual library of scholarly Internet resources (predominantly Web sites) built using the iVia Virtual Library Software<sup>2</sup> [27]. It contains a core set of expert-created metadata records, which are augmented by sets of imported expert-created records created by collaborating institutions, and by a large set of

---

<sup>1</sup><http://infomine.ucr.edu>

<sup>2</sup><http://infomine.ucr.edu/iVia>

secondary records created by automatic processes without direct human intervention.

At the time of writing, there were approximately 100,000 automatically-created records in INFOMINE, comprising around 50% of the collection, and hundreds of thousands more in supporting collections that are used internally by INFOMINE or exported to collaborators. Most of the resources are selected by INFOMINE's two Web crawlers, the Virtual library Crawler and the Automatic Focused Crawler, which continuously crawl the Internet identifying scholarly resources [27].

## 2.1 Approaches to Assignment

There is an abundance of prior work on metadata assignment, but most of it focuses on a single field (or a few related fields), whereas this paper takes a broader view, describing and evaluating a range of assignment methods.

Most metadata assignment techniques follow one of two approaches: extraction or classification. Extraction techniques assign values drawn from the text of the document. Keyphrase extraction algorithms, for example, work by parsing the text of a document and choosing a subset of its words and phrases to assign [7][29], and "harvesting" tools extract values from HTML Meta tags. Extraction is most appropriate for "uncontrolled" fields (i.e. those that do not have a controlled vocabulary) like Title, Description, Creator and Keywords.

Classification approaches assign metadata values from a controlled vocabulary. Most classification techniques work in two phases: training and assignment. In the training phase, a training set of documents labeled with values from the vocabulary is built, and then used by a classification algorithm to learn the relationship between the documents' features and labels, and to store this knowledge in a model. In the assignment phase, this model is used to classify new documents, and assign them the most appropriate labels. Classification techniques are often used to assign metadata to fields like Language, LCC and LCSH.

There are some programs for extracting a range of metadata values, including the DC-dot program<sup>3</sup>, which extracts Dublin Core metadata from the author-supplied Meta tags in HTML documents, and MetaExtract, which assigns (and evaluates) a variety of metadata based on natural language processing techniques [33]. Another application, described by Han *et al.* [12], extracts 15 types of fielded metadata from consistently-structured research papers by training a classifier to recognize the different metadata types. Specialized assignment methods will be described in Section 4 in connection with the related iVia tool.

## 2.2 Approaches to evaluation

There are two broad approaches to evaluating metadata assignment: automatic evaluation by computer programs, and human evaluation by subject domain experts. These are forms of quantitative and qualitative evaluation respectively.

An automatic evaluation requires a set of documents whose expert-assigned metadata values are known. Metadata values are automatically assigned to the documents, and then the degree of similarity between the automatically-assigned values and the expert-assigned values is measured. This is a natural adaptation of classical machine learning and statistical evaluation to a new domain—document metadata—using standard performance measures from machine learning and information retrieval.

A human evaluation involves taking a set of documents, assigning them metadata, and then having a group of subject domain (or cataloging) experts rate the appropriateness of the metadata assigned [1][16][17][18][33]. Human evaluations have many advantages. They can be applied to a wide range of documents, they measure the actual usefulness of the result (rather than measuring the similarity of the result to some standard), and human assessors can allow for complexities like near misses that may be overlooked by automatic evaluations. However, human evaluations are expensive in terms of time and resources, so automatic evaluations are more common.

Most work on automatic metadata evaluation considers specific fields. Larson's study of LCC assignment was possibly the first automatic evaluation of metadata assignment [21]. Similar approaches were used by Turney [29][30] and Frank *et al.* [7] to evaluate different Keyphrase extraction algorithms on technical documents by comparing their assignments to those made by the documents' authors. Both Dolin [5] and Godby *et al.* [10] report automatic evaluations of LCC assignment based on library catalogs. Frank and Paynter [9] also evaluated LCC assignment against a library catalog, and consider issues such as measuring partial matches in subject hierarchies. There is a related and growing body of work on text classification, for example using Medical Subject Headings (e.g. [28]) and Internet subject hierarchies like DMOZ and Yahoo (e.g. [3][26]), which tend to follow the evaluation methods of the machine learning community. Description metadata, in the form of automatic summaries, have been evaluated with N-gram based metrics [24].

Automatic evaluation has several drawbacks. It is limited to sets of documents where the metadata values are known (which may require significant preparatory work, though test sets are usually chosen because such metadata is already present). It is based on the assumption that the known metadata really is good metadata, and can have difficulty expressing complexities such as "nearly correct" assignments, or cases where there are alternative, but unknown, correct assignments. Despite these flaws, the speed and ease of performing automatic evaluations make it a valuable tool for refining assignment techniques when a suitable test collection is available.

## 3. AUTOMATIC METADATA EVALUATION

The iVia software includes an metadata assignment evaluation program, which selects a large number of metadata records with expert-created metadata, assigns them metadata based on their URL field, and then measures the quality of the assigned metadata by comparing it to the expert-assigned metadata.

Metadata quality is often difficult to define, so we adopt the practical view expressed by Guy *et al.*: "high quality metadata supports the functional requirements of the system it is designed to support" [11], and emphasize different quality measures depending on the purpose of each metadata field.

### 3.1 Test data

Test data for the evaluation program are chosen from an iVia database. The program allows control over the criteria by which records are chosen, the number of records chosen, the order in which records are selected, and the metadata fields that are used in the evaluation.

The configuration used at INFOMINE, and in the experiments below, is to evaluate metadata on 1000 expert-created records

---

<sup>3</sup><http://www.ukoln.ac.uk/metadata/dcdot/>

imported from INFOMINE, sorted in order of most recent modification (newest first). We choose the most-recently-modified records that meet our selection criteria (as opposed to a random selection) because these are recently reviewed and therefore likely to contain the highest quality metadata, and so that our optimizations accommodate changing trends in INFOMINE indexing practice. The records are further restricted to resources that can be downloaded (pages that are stale, fee-based, or inaccessible due to robots.txt files, are not used) and are HTML documents.

Automatic evaluation depends on expert-assigned metadata to calculate performance statistics. Our primary test dataset is drawn from the collection, where every expert-created INFOMINE record has a set of required fields, including URL, Title, Creator, Description, Keyphrase, LCSH and INFOMINE Category, and a set of optional fields, including LCC, Resource Type, Audience level and INFOMINE Subject Discipline, which are assigned at the discretion of the editor. Records are also assigned a Language (always English) and Media Type (also known as MIME Type or Format) automatically. This evaluation is limited to INFOMINE's required fields: Title, Creator, Description, Keyphrases, LCSH, and INFOMINE Categories. More extensive evaluations can be performed by importing new collections into iVia using the MARC or OAI-PMH importers, but this is beyond the scope of the present paper.

### 3.2 Evaluation statistics

Different metadata fields have different characteristics. For example, some fields are only assigned a single metadata value, in which case accuracy is a good performance measure, while others are assigned multiple values, and precision and recall are more informative. Some fields are even more complex: accuracy is not a useful way to measure Description assignment, for example, and fields with complex metadata types like LCC and LCSH may be “partly” or “nearly” correct. Consequently, the iVia evaluations use a range of different performance measures.

The most basic statistic is *exact match accuracy*, which is the proportion of times the automatic assignment for a field exactly matches the expert's assignment for a field. Matches apply to all the values in multiple-value fields. Because of this strict interpretation, exact match accuracy is generally only useful in controlled fields where one value is assigned, such as Language or Media Type.

In multiple-value fields like Keyphrases, an assignment will not be considered an “exact match” unless all the correct values are assigned to a resource, and no incorrect values are assigned. It does not account for cases where some, but not all, of the values are correctly assigned. To remedy this shortcoming, we split each multiple-value field into its individual *subfields*, and then measure performance using two statistics: *subfield precision* and *subfield recall*. These statistics are defined as follows:

$$\text{subfield precision} = \frac{\text{number of correct assignments}}{\text{total number of assignments}}$$
$$\text{subfield recall} = \frac{\text{number of correct assignments}}{\text{total number of expert-assigned values}}$$

In other words, subfield precision is the proportion of the assignments that are correct, and subfield recall is the proportion of the expert-assigned values which are matched by the assigner.

Exact match accuracy, subfield precision, and subfield recall are good for measuring the performance of controlled vocabulary terms, where descriptors always take on known forms (or simple forms), but less useful for measuring fields with uncontrolled text

values, like Description and Title. Even a very good tool for generating textual summaries will rarely generate exactly the same description as a human expert. To evaluate textual fields, we have introduced two new metrics: *content-word precision* and *content-word recall*. Both metrics compare two passages of text by analyzing the set of unique content words (i.e. words that are not common stopwords like *of* and *the*) present in each passage (ignoring case). Thus the content-word precision is the proportion of content words in the assigned value that are also present in the expert-assigned value; and content-word recall is the proportion of the content words from the expert-assigned value that also appear in the automatically-assigned value.

A problem with some fields—particularly Keyphrases—is that content words can appear in plural or singular form in both the expert-assigned and automatically-assigned metadata, and this can result in nearly-correct matches, like *library* and *libraries*, being ignored. A simple way to counter this problem is to “stem” the content words before calculating precision and recall. We use an implementation of Lovins' stemmer [25], and call the resulting statistics the *stemmed content-word precision* and *stemmed content-word recall*.

### 3.3 Workflow

The evaluation tool is used to guide the development of the iVia metadata assignment tools, not to make a definitive statement about the quality of the metadata assigned. Many of the metrics are not useful in isolation: instead, their role is to demonstrate improvement in metadata assignment processes over time. In practical terms, this means that we use automatic evaluation to test whether a change has beneficial or detrimental effects.

To give a concrete example, it was recently noted that INFOMINE contains thousands of automatically-created records whose Title that starts with the prefix *Welcome to* (examples range from “Welcome to AADP” to “Welcome to Zoo Atlanta”). In many cases, the prefix is superfluous, but in others, it is required (for example, in reference to the motion picture “Welcome to Sarajevo”). The Title assignment tool was updated to remove the prefix from all records, and an evaluation showed the benefits of the change significantly outweighed the costs.

## 4. AUTOMATIC METADATA ASSIGNMENT

Automatically-created records in iVia are assigned metadata by a set of modules we collectively call the *iVia metadata assignment tools*. The tools use a wide range of extraction and classification techniques, and can be used individually or in combination.

The tools are part of the iVia Metadata Assignment library (libiViaMetadata). They are usually applied to HTML pages downloaded from the Internet, though they can also be applied to local documents, and to documents in other formats such as PDF files. In the evaluations below, the library is responsible for downloading the Web pages and ensuring they are HTML documents. When HTML framesets are encountered, their constituent frames are composed into a single document that approximates the content that is displayed by a Web browser.

This section describes each of the metadata assignment tools in iVia. In each case, the metadata field is described in a general sense, and then as it is applied in INFOMINE, and the appropriate evaluation measures are identified. The assignment process (or processes) are then described, and evaluated with the evaluation tool. Finally, any relevant related work is discussed.

## 4.1 Title Assignment

Titles are short, uncontrolled text passages that appear at the beginning of many types of document. They are usually chosen to identify the document and summarize its content. Both roles are important, though the latter is most useful when assessing an unknown document, for example in a list of search results. Most applications require a single Title value for each resource.

Titles are particularly important in INFOMINE searches because the standard ranking system assigns a greater importance to Title metadata than any other field. INFOMINE makes heavy use of Title metadata when presenting documents in search result lists, browse pages, and other applications.

### 4.1.1 Evaluation Measures

Both recall and precision are important in Title assignment, and as Titles contain uncontrolled text values, content-word-based statistics are very useful. Broadly speaking, a high recall score indicates the important words in the Title are identified correctly, while high precision suggests we are not assigning incorrect words that summarize unimportant parts of the document and skew search results. By convention, Titles are often short and predictable, so exact matches are frequently possible. For these reasons, the primary measures in our evaluations will be content word precision, content word recall and exact match accuracy.

### 4.1.2 Assignment Process

Title assignment is an extraction process: Title values are simply read from appropriate parts of the HTML document, such as the *Title* tag. Although it might appear that this method will yield a 100% success rate, many HTML authors supply no Title, or poor values that are corrected before being included in INFOMINE.

A list of potential titles is built up by extracting text from the following sections of the HTML document, in order:

1. The content of any *Meta* tag whose *name* is *title* or *dc:title*.
2. The *Title* tag.
3. All *H1* tags.
4. The sequence of words in the first 50 letters of body text.

The initial list is post-processed to remove duplicate entries, blacklist undesirable values (e.g. *Homepage*, *Untitled Document*), and remove unwanted prefixes (e.g. *Welcome to*, *Homepage of*) while preserving the order of the list. The values remaining in the list are assumed to be in order of decreasing quality, so that when a single Title is required, the first is used.

### 4.1.3 Evaluation

Table 1 summarizes our evaluation of the first Title assigned by the process above. The columns list the method name; the number of times a Title was assigned by that method; the exact match accuracy (EMA); the content word precision (CWP) and recall (CWR); and the length in words. Each row shows the result of using a different method to extract Titles, starting with the current method (described above), which assigns the exact Title in about 23% of cases, and has content-word precision of 62% and recall of 64%. The assigned metadata has an average length of 5.5 words, the same as the expert-assigned metadata.

Rows 2 to 5 show the result of using only one of the sources of Title metadata at a time (with standard post-processing). Document Meta tags (row 2), though rare, are less precise than Title tags (row 3), so arguably Title tags should appear before Meta tags in the list of preferred sources above. An automatic evaluation shows that this change does not affect the overall performance, and a side-by-side comparison reveals that the Meta

Table 1: Title assignment evaluation results

	Method	Tries	EMA	CWP	CWR	Length
1	Current	1000	0.2290	0.6200	0.6410	5.5
2	Meta only	21	0.0050	0.4783	0.0132	7.2
3	Title only	992	0.2340	0.6295	0.6434	5.5
4	H1 only	165	0.0350	0.7289	0.0786	3.6
5	Text only	919	0.0000	0.2390	0.2931	8.0

Table 2: Creator assignment evaluation results

	Method	Tries	SFP	SFR	CWP	CWR
1	Current	151	0.0724	0.0049	0.4125	0.0425

tags are usually either identical to Title tags, or very similar but longer and more descriptive. Although the last two sources (rows 4, 5) perform poorly compared to the Meta and Title tags, they are useful for PDF and other non-HTML documents.

### 4.1.4 Related work

Title assignment is extremely simple, and similar techniques are used by almost every Internet search tool—though we suspect these algorithms are usually considered too simple to require evaluation. For example, DC-dot explicitly extracts Title Metadata from Title and Meta tags, and iVia's final method (extracting words from the beginning of the text) follows an implementation in the Greenstone Digital Library Software [31].

## 4.2 Creator Assignment

Creator metadata identifies the person or organization primarily responsible for the creation of a resource, and is often called Author metadata. Documents can have one or more creators, though additional authors may be better described as contributors or publishers, depending on their specific role. The creator-contributor-publisher trichotomy is drawn from the Dublin Core Metadata Element Set [6]. INFOMINE sometimes contains alternate forms of organization names. For example, some records have *United States Forest Service*, *US Forest Service* and *USFS* assigned as Creators.

### 4.2.1 Evaluation Measures

Creator is a multiple value field whose values have an uncontrolled but limited vocabulary. Values are not always assigned in a consistent format, and a personal name could appear as either *John Smith* or *Smith, John*. For this reason, content-word precision and recall are appropriate measures.

### 4.2.2 Assignment Process

Creator assignment is the simplest of the metadata assignment tools. Metadata is extracted from the content attribute of any Meta tag whose name is *creator* or *dc:creator*. A blacklist is then used to eliminate any unwanted terms, though the blacklist is very short, containing only the value *none* and the prefix *unknown*. Contributor and Publisher metadata are assigned by the same method (with different Meta tags, but the same blacklists).

### 4.2.3 Evaluation

The Creator metadata evaluation is reported in Table 2. Because few documents supply Creator metadata, an assignment was made

**Table 3: Keyphrase assignment evaluation results**

	Method	Tries	Number	SFP	SFR	CWP	CWR	SCWP	SCWR
1	Current	987	9.21	0.1402	0.0682	0.2807	0.1677	0.3093	0.1938
2	Meta tags only	457	3.65	0.2172	0.0422	0.3165	0.0774	0.3475	0.0886
3	PhraseRate only	987	9.11	0.0901	0.0437	0.2626	0.1555	0.2916	0.1818

in only 151 of the 1000 cases, and recall is low. For the metadata assigned, content-word precision is high, at 41%, while subfield precision is low, at 7%, suggesting that the metadata provided by the authors tends to be inconsistently formatted.

#### 4.2.4 Related work

As in the case of Title metadata, Creator extraction is very simple and seldom warrants explanation or evaluation. DC-dot also extracts Creator, Contributor and Publisher metadata from Meta tags. In addition, DC-dot can attempt to assign Publisher metadata to an Internet resource by looking up the owner of the domain name of the Web site hosting the resource. We believe this process shows promise, but will require human refinement to be sufficiently accurate for our applications. Another promising approach is to use name authority files to support Creator metadata creation, as in cataloging and other metadata tools.

### 4.3 Keyphrase Assignment

Keyphrases are a set of short, uncontrolled phrases that describe the content of a document. Author-assigned keyphrases in technical documents typically consist of five to fifteen complementary phrases of between one and four words.

Keyphrases in the INFOMINE test set have been assigned by librarians, and have different characteristics. Each document has been assigned an average of 19.0 keyphrases, often including plural forms and “semi-controlled” values. A semi-controlled value is one agreed upon by catalogers to describe a type of resource, but not (yet) formally codified as part of a controlled vocabulary. Examples include *Virtual Library* and *E-text*.

Keyphrases are sometimes used to present lists of topics in browsing interfaces, and to display documents, so the INFOMINE editors have explicitly set the maximum number of assigned phrases to ten, the maximum phrase length to five words, and the minimum phrase length to two words.

#### 4.3.1 Evaluation measures

Keyphrase metadata is used to summarize the important content of a document in retrieval and display. Precision is important because the Keyphrase field is relatively highly weighted in INFOMINE, and because the potential of inaccurate keyphrases to misrepresent the content of the document is exaggerated by the fact that relatively few phrases are used to represent entire documents.<sup>4</sup> Recall is also important, as it approximates the degree of coverage of the document subject matter that has been attained. Our evaluation will therefore focus on subfield precision and recall, and content-word precision and recall. Since the test data often contains plural forms, the stemmed content-word recall is the best measure of coverage.

<sup>4</sup> However, our assignment algorithm was designed for interactive use, and explicitly targets recall, noting that “Coverage of the top keywords is more important than precision since there is a human selector” [13].

#### 4.3.2 Assignment Process

The iVia Keyphrase assignment module combines keyphrases from two complementary sources, then ranks and post-processes the results. The first source is any HTML document Meta tag named *keyword* or *keywords*. Because this source is infrequently available, and notoriously abused by some authors (who use the field to garner search engine rankings, not to describe their document) it cannot be used in isolation.

The second source of phrase data is iVia's PhraseRate keyphrase assignment engine [13]. PhraseRate is explicitly designed to extract phrases of two to five words from pages that are “well written and subject oriented”, and to require no training phase and no knowledge of word distributions. PhraseRate works by assigning a document-wide score to each word that appears in the document based on the HTML markup, distance from the beginning of the paper, and capitalization of each occurrence. It then extracts every well-formed phrase of two to five words and assigns each a score based on its constituent word scores. Finally, the phrases are post-processed to eliminate any phrase that is a sub-phrase of a higher-scoring phrase, the phrase scores are normalized to the range  $[0,1]$  and the highest-scoring remaining phrases are assigned. For a detailed description, see [13].

Once extracted, the candidate keyphrases from the document Meta tags and from PhraseRate are combined in a single list, and assigned a score. Phrases are initially awarded a score of 1.0 if they appear in the Meta tags, and 0 if they do not. The scores of the phrases extracted by PhraseRate are then incremented by their PhraseRate score (which is less than 1.0). A small bias is introduced towards longer phrases to break ties: each phrase's score is incremented by its length (in letters) X 0.000001. Any blacklisted phrases are removed. The remaining phrases are then sorted, and the ten highest-scoring phrases are returned.

#### 4.3.3 Evaluation

Table 3 shows the result of the keyphrase assignment evaluation. The columns show the method used; the number of times Keyphrase assignment was attempted; the average number of phrases assigned; the subfield precision (SFP) and recall (SFR); the content-word precision (CWP) and recall (CWR); and the stemmed content-word precision (SCWP) and recall (SCWR). The first row shows the performance of the current keyphrase assignment process, described above.

Row 2 of Table 2 shows the performance when only HTML Meta tags are used for assignment, and row 3 shows the effects of using only PhraseRate. Meta tags tend to have higher quality metadata, but it is frequently unavailable, and when it is available fewer than four phrases are assigned (on average), resulting in low recall. PhraseRate makes assignments in 99% of cases, and assigns more than nine values (on average) to each document. It has higher recall, but lower precision.

#### 4.3.4 Related work

Most Keyphrase extraction algorithms follow the same pattern: a set of candidate phrases are extracted from a document, the candidates are ranked, and the best candidates are assigned. They are evaluated by comparison to the “keywords” assigned to technical reports by their authors. Early work by Turney used GenEx, a hybrid genetic algorithm, to learn the optimal parameters for the Extractor keyphrase extraction tool [29]. The New Zealand Digital Library Project developed a very simple algorithm, Kea [7], based on a Naive Bayes classifier [32], that offers similar performance and has been extended in several ways [7][30]. Barker and Cornacchia use natural language processing techniques to extract noun phrase heads for use as keyphrases, and use a human evaluation to show their system performed approximately as well as Extractor [1]. A series of human evaluations by Jones and Paynter found that Kea and Extractor tend to assign good phrases, but that human experts assign much better phrases, a result that supports their use in automatic evaluations [16][17][18].

### 4.4 Description Assignment

Description metadata provides a concise, textual representation of the intellectual content of a work. In most applications this means one or two paragraphs of text that describe the subject matter of the work, such as a summary or abstract, though it can have other forms, such as a table of contents.

Description metadata is often used to summarize a document. A single, textual passage is used to give a concise but expressive summary of a work, usually in conjunction with other metadata, such as a Title and list of authors. In this case, only a single Description element is required. Descriptions are also valuable fields for supporting search, since they provide a relatively large amount of text (compared to other fields), yet are shorter (and have more consistent lengths) than the full text they summarize.

INFOMINE has a generic Description field that is required in all records, and contains a summary of the resource suitable for display. Additional Descriptions may be supplied for use in special applications, and while these are indexed to support user searches, they are not considered in this evaluation.

#### 4.4.1 Evaluation Measures

INFOMINE's Description field acts as a summary when records are displayed, and also supports user searches. In each case, it is desirable that all the topics covered in the document are represented in the summary, so recall is an important measure. A good summary will not mislead the reader by referring to unimportant material, so precision is also important. Because only one Description is assigned, and it is a free-text field, our primary measures are content word recall and precision (both stemmed and unstemmed). We also monitor the average length of the assigned metadata to ensure it is suitable for presentation.

#### 4.4.2 Assignment Process

The iVia Description assignment process is based on two sources: HTML Meta tags and a text summarization algorithm. The first step is to check for Metadata tags named *description* or *dc:description*, and if either is present, they are used as the description. If that fails, a text summarization program is used to extract a summary.

The summary is generated by *AutoAnnotator* [19]. This method breaks the HTML document down into sections reflecting its structure, and then finds the single paragraph of text that best represents the content of the work. The algorithm is initialized

**Table 4: Description assignment evaluation result**

	Method	Tries	Len.	CWP	CWR	SCWP	SCWR
1	Current	992	36.2	0.2977	0.1891	0.3215	0.2072
2	Meta tags only	464	20.5	0.3907	0.0776	0.4253	0.0865
3	AutoAnnotator	992	51.3	0.2284	0.1938	0.2471	0.2118

with an HTML document, and a set of words that are “important” to the document. In iVia, these important words are the content words appearing in the extracted Title and Keyphrase metadata.

*AutoAnnotator* is based on sentence and paragraph scoring. First, the document is read, and split into textual divisions, which are then split into paragraphs, which are split into sentences, which are finally split into words. Each word is then assigned a score, which is based on several factors, including whether the word is an “important” word, whether it is a stopword or content word, and whether it appears as plain text or decorated by HTML markup (e.g. heading tags, bold tags). Once the words are scored, they are used to calculate a combined score for the sentence, paragraph, and textual division they occur in. These scores are then modified to account for the position of the text in the document, as text that appears early in the document is more likely to contain a useful summary. Finally, the highest-scoring text division is found, and the highest-scoring paragraph within it is returned. If this strategy fails, a set of contiguous high-scoring sentences are returned instead. For a full description, see [19].

#### 4.4.3 Evaluation

Table 4 shows the result of a recent evaluation. The columns list the assignment method used; the number of times a description was suggested; the average length of the description in words; the content-word precision (CWP) and recall (CWR); and the stemmed content-word precision (SCWP) and recall (SCWR).

The first row of Table 4 contains the results of the current method, described above. A description was suggested on 992 of the 1000 attempts, and the description had an average length of 36.2 words—much shorter than the expert-created records, which had an average length of 59.0 words. The stemmed results are uniformly better than the unstemmed equivalents, suggesting that the terminology used in the assigned descriptions is similar, but not identical, to the terminology used by INFOMINE's Editors. Rows 2 and 3 show the performance of assignment by Meta tags only, and the performance of the *AutoAnnotator* only.

Interestingly, the 992 assignments included three exact matches (two from Meta tags, one from *AutoAnnotator*), reflecting the occasional practice of creating a description by quoting directly from the resource.

#### 4.4.4 Related work

There is an extensive body of literature on Description extraction, and many of the ideas in *AutoAnnotator* have been previously reported: sentence-scoring and paragraph-scoring are established methods, and the use of extracted keyphrases as the basis of scoring has been described elsewhere [15]. Our implementation is informed by these precedents, and is robust and fast.

Several commercial summarization products are currently available. Another implementation, the Open Text Summarizer,<sup>5</sup>

<sup>5</sup><http://libots.sourceforge.net/>

**Table 5: LCSH assignment evaluation results**

	Method	Tries	SFP	SFR	LHP	LHR	CWP	CWR
1	Current	683 of 1000	0.1894	0.2110	0.3136	0.3259	0.2505	0.2821
2	Chan [4]	100 of 100	0.3684	0.3271	0.6105	0.5421		

is a Free Software tool that is similar to *AutoAnnotator*, but works on plain text (not HTML), and uses a different set of important words for sentence scoring (the document's content words, each weighted according to its frequency in the text).

Lin and Hovy describe a set of evaluation metrics for Description metadata based on N-gram co-occurrence statistics, and demonstrate they are highly correlated with human evaluations [24]. We intend to apply their work in iVia, for Description assignment, and for other metadata fields.

## 4.5 LCSH Assignment

The LCSH are, as the name suggests, a set of over 200,000 Subject Headings created and maintained by the Library of Congress [23]. They consist of short, descriptive phrases, which may be modified by combining a Head term with one or more subdivisions: for example, the Head term *History* might have two subdivisions added to produce *History -- United States -- 19th Century*. Usually, a set of LCSH will be assigned to a document. The Library of Congress assignment rules suggest that the first LCSH represent the primary topic, and subsequent LCSH represent less-important topics, but indexers do not always agree on a document's topics [4], and the practice is not used in INFOMINE, so it is ignored in this analysis.

The primary use of LCSH, both in library catalogs and in INFOMINE, is for retrieval, as experienced librarians can perform very discriminating searches based on their knowledge of LCSH. They are also used for detailing the content of a document in a concise form when records are displayed.

### 4.5.1 Evaluation Measures

LCSH are a controlled vocabulary, and multiple values are assigned, so subfield precision and recall are the obvious measures of performance. However, the number of LCSH is very large, and the number of potential combinations even larger, so different sources will often assign different LCSH—even human catalogers will frequently disagree [4]—and a more complex metric that allows for near misses is appropriate. Therefore we consider the content word precision and recall, and the LCSH Head precision and recall. The latter measure is the same as subfield precision and recall, but uses unique LCSH Heads as subfields (instead of the full LCSH).

### 4.5.2 Assignment Process

iVia assigns LCSH using a classification process that is conceptually related to k-nearest-neighbor or locally-weighted learning [8], but which has been extended to handle examples that routinely belong to several categories, and that belong to classes with very large vocabularies.

The assignment process depends on a supply of expert-assigned training data, in our case the INFOMINE collection. LCSH are assigned to a new document in two stages: first, the set of documents that are the most similar to the new document are discovered; and second, the LCSH that are assigned to the similar documents are retrieved, and the most popular are assigned to the new document. Each stage has several complexities.

In the first stage, the goal is to find the set of  $N$  most-similar documents (in practice,  $N$  is set to 15). This set is approximated

by exploiting iVia's strengths. The Keyphrase assignment module is used to extract a set of keyphrases from the document, and these are formed into an disjunctive query for iVia's search engine [27]. For example, if the keyphrases *Africa*, *Sahara desert*, and *sand dunes* are assigned, they will be combined in the query *africa OR (sahara AND desert) OR (sand AND dunes)*. The query is used to search the Title, Keyphrases, Description and full text fields of the iVia database (results are sorted using the standard iVia ranking system [27]) and to retrieve the  $3 \times N$  best results. Next, a "similarity score" is calculated for each of the results, by using the cosine measure to compare them to the original document, resulting in a number between 0 and 1. The  $N$  most similar documents are then chosen.

The second stage proceeds by retrieving the LCSH metadata from each of the similar records and assigning a score to each full LCSH that occurs. The score is calculated by adding together the similarity scores for each of the documents that the LCSH appears in. This can be thought of as a voting process, where each of the similar documents "vote" for their LCSH, and their votes are weighted according to how similar each LCSH is to the target document. LCSH that appear in only one of the documents are eliminated, and the six<sup>6</sup> remaining LCSH with the highest scores are assigned.

The LCSH assignment algorithm reduces the complexity of the LCSH problem using a form of locally-weighted learning. Instead of trying to assign six descriptors from the hundreds of thousands of possible values, we instead narrow the number of possible descriptors to the small subset that occur in the 15 most-similar documents. This requires considerable computation at classification time (as opposed to during training), and has the potential to be very slow. However, iVia's search functions have been optimized for speed and relevance over a period of several years, so the LCSH assignment speed is still acceptable.

### 4.5.3 Evaluation

Table 5 shows an evaluation of the LCSH assignment algorithm. The columns identify the method used; the number of times the method was able to make a prediction; the subfield precision (SFP) and recall (SFR); the LCSH Head precision (LHP) and recall (LHR); and the content-word precision (CWP) and recall (CWR). The data in row 1 shows the current algorithm, described above. As in the case of uncontrolled keyphrases, the precision and recall are low, and disguise many near-misses.

LCSH assignment is particularly difficult for machine learning systems because good results are dependent on good training data, and there are so many different LCSH that it is difficult to find training data relating to all of them. In terms of LCSH assignment, this means that the algorithm can only assign LCSH accurately when there is already a set of similar documents that have been assigned appropriate LCSH by human experts. In INFOMINE, this means that when no similar documents exist, the search engine tends to return a set of documents that have few or coincidental similarities to the target document, which biases the

<sup>6</sup>This constant was selected by the INFOMINE Editors.

**Table 6: INFOMINE Category assignment evaluation results**

	Method	Tries	Number	SFP	SFR	EMA	Date
1	Current	981	1.5	0.8195	0.8854	0.6960	2005-01-28
2	No GovPub rule	980	1.5	0.8126	0.8826	0.6920	2005-01-28
3	Naive Bayes variant	382	1	0.7357	0.5249	0.4467	2005-01-06
4	Old kNN variant	676	1	0.7204	0.6040	n/a	2004-11-22

assignment towards very common LCSH like *California*, *Periodicals*, and *Web Sites – Directories*.

#### 4.5.4 Related work

Although there is little reported work on automatic LCSH assignment, Chan describes a similar evaluation in her study of human indexer consistency [4]. Using a set of library catalog records that had been assigned LCSH both by human experts at the Library of Congress, and by human experts at other libraries, she evaluated how closely the other libraries' assignments matched the Library of Congress' assignments. She reports an exact match accuracy of 15%: only 15 records out of 100 were assigned the same sets of LCSH by both groups of humans (and six of these were assigned zero values). As in the present research, Chan found this statistic unsatisfactory on its own, and explored more descriptive measures. Some other statistics, calculated from [4], are shown in row 2 of Table 5. Chan's expert-assigned metadata had a subfield precision of 37% and a subfield recall of 33%, which is almost twice the performance reported by the automatic system.

#### 4.5.5 Future work

The current LCSH assignment scheme has several limitations. First, it does not have enough knowledge of the structure of LCSH descriptors to perform useful generalizations about their meaning, which causes it to overlook relationships between similar LCSH. For example, two LCSH with the same Head term but different subdivisions are related, but the LCSH assigner treats them as distinct topics (which can cause it to overlook the Head term entirely). Second, LCSH assignment is dependent on having training data available at classification time, so can only be applied in an iVia installation where suitable training data is available. Finally, the algorithm itself is relatively simple. We are replacing it with a system that discovers structure within the topical LCSH using expert knowledge and automatic clustering, and then induces a hierarchical classifier similar to the one proposed for LCC assignment (Section 4.7.2).

## 4.6 INFOMINE Category Assignment

The INFOMINE Categories are used to divide the INFOMINE collection into nine sub-collections, each of which is managed by a different editor, and which can be used to filter searches and restrict browsing to general topic areas. The set of collections does change (albeit very infrequently) and currently includes seven topic-based collections (*Biological*, *Agricultural and Medical Sciences*, *Business and Economics*, *Humanities*, *Visual and Performing Arts*, etc), one format-based collection (*Electronic Journals*), and one publisher-based collection (*Government Publications*). Every expert-created record in INFOMINE is assigned to one or more of these categories.

### 4.6.1 Evaluation Measures

The INFOMINE Categories are a simple example of a controlled metadata element which allows multiple assignments.

Consequently, the two primary evaluation metrics are subfield precision and subfield recall. Because the number of possible values is so small, and the quality and quantity of training data is high, we can frequently assign the entire set of records, so it is also useful to measure the exact match accuracy.

### 4.6.2 Assignment Process

iVia assigns INFOMINE Categories using a set of binary classifiers, each of which is responsible for assigning a single category based on the text of a document. Two steps are involved: training and classification. The training process is run weekly as part of iVia's automated maintenance scripts, and takes several hours,<sup>7</sup> while the classification step is invoked whenever an assignment is requested, and is almost instantaneous.

The training step builds a set of *category classifiers*, each of which is a probabilistic binary classifier that classifies a new example as either belonging to, or not belonging to, a particular category. The training data are INFOMINE records that describe Web pages (and whose full text is available). Each category classifier builds its own training set, which must have at least 1000 positive and negative training examples, otherwise the category is skipped (the smallest INFOMINE Category, *Cultural Diversity*, is routinely skipped). The training data is then reduced to limit both positive and negative classes to 10,000 training examples, to ensure that the ratio of negative to positive examples (and vice-versa) is no greater than 3:1, and to limit the number of features to 50,000 words using Chi-squared feature selection. Once the training set is build, the category classifier is trained on the data. In the current implementation, the category classifiers are binary Logistic Regression classifiers [20].

The classification step is invoked to assign INFOMINE Category metadata to a new resource. Each of the category classifiers is loaded and applied to the text of the new document, and their individual predictions are combined to assign a set of categories. The procedure for combining individual assignments is:

1. Any category assigned with confidence greater than 0.95 is automatically assigned; but
2. If no categories are assigned by step 1, then the category (or categories) with the highest probability is assigned, unless its probability is less than 0.75, in which case no assignment is made.

There is one exception to this process, specified by the INFOMINE Editors. If a resource is hosted on a server whose domain name ends in *.gov*, then it is always assigned the category *Government Publications*, regardless of the output of the relevant category classifier.

<sup>7</sup> Training takes approximately 12 hours running in the background on an 2000 Megahertz workstation under moderate load with 19,000 previously-cached training examples.

### 4.6.3 Evaluation

Table 6 shows the results of several evaluations of INFOMINE Category assignment. The columns identify the method being used for assignment, the number of times an assignment was attempted, the subfield precision (SFP), subfield recall (SFR) and exact match accuracy (EMA). The rightmost column is the date that the evaluation was performed: the same process was used to select the training and test data for each evaluation, but different documents were used in the earlier evaluations, reflecting the state of the INFOMINE database on those dates.

The first row shows the current method, described above. On average 1.5 categories were assigned to each record (the same as the human experts) and precision was 84% and recall 89%. Row 2 shows that removing the special rule for assigning *Government Publications* to *.gov* Web sites has almost no effect.

The next rows show the equivalent statistics for two other assignment methods. Row 2 shows the result of replacing the set of Logistic Regression classifiers with a set of Naive Bayes classifiers [32], and shows that Logistic Regression outperforms Naive Bayes by every measure: more assignments are attempted, and precision, recall, and exact match accuracy are all much greater. Row 3 shows the results for an older assignment algorithm, which operated in the same way as the LCSH assignment algorithm (Section 4.5.2). This algorithm is also inferior to the current algorithm by all measures, though it is arguably preferable to the Naive-Bayes-based method (row 2), as it makes predictions in more cases yet has comparable precision and superior recall.

### 4.6.4 Related work

There is a large and growing body of work on text classification. An obvious alternative to the scheme described here is to use a multi-class Logistic Regression classifier. In fact, this has been implemented in iVia, but is not current used because the training process takes too long.

The INFOMINE Category assignment algorithm requires that each binary classifier is probabilistic (i.e. it must estimate the probability that its result is correct), which limits the classification schemes that can be used. The “standard” probabilistic classifier is Naive Bayes, which is versatile and fast, but often exhibits poor performance relative to modern classifiers (as can be seen in Table 6). Many modern alternatives, such as the Support Vector Machine algorithm used by INFOMINE for LCC assignment (described below) are either too slow for our application, or cannot produce the necessary probability estimates. Logistic Regression classification was chosen for use in iVia after careful deliberation because it promised accuracy comparable to Support Vector Machines, speed comparable to Naive Bayes, and probabilistic assignment. Happily, that promise has been largely fulfilled. Since this project began, new work on Support Vector Machines has emerged that attempts to address the issues of speed and probabilistic output.

## 4.7 Other Metadata Assignment Tools

iVia commonly assigns Metadata to the following fields for which no test data is available in the INFOMINE dataset.

### 4.7.1 Language Assignment

INFOMINE is a purely English-language resource, and any automatically-discovered records must also be predominantly in English. iVia assigns Language metadata based on two sources. First, if HTTP headers are available and contain a *Content-Language* field, the value of that field is used. If that fails, iVia assigns a language based on the content of the document with an implementation of the N-gram-based algorithm outlined by

Cavnar and Trenkle [2]. The N-gram method is widely used, and widely considered reliable. The authors report accuracy of over 97% if sufficient training data is available [2].

### 4.7.2 LCC Assignment

The Library of Congress Classification is a set of hierarchical descriptors that are usually used to identify the main topic of a work [22]. INFOMINE's LCSHtoLCC tool is based on the Support Vector Machine algorithm, and uses LCSH Head terms to assign LCC [9]. It represents a significant advance on previous approaches based on information retrieval techniques [5][10][21]. LCSHtoLCC has higher accuracy, covers a larger portion of the LCC topic space, and provides a more detailed evaluation [9].

The LCC assignment algorithm is elegant and effective, but the implementation has several drawbacks: it is not fully integrated into iVia, it is slow to train, it cannot assign multiple classes, it requires that LCSH are already assigned, and it does not assign a probability with its predictions. We are addressing these problems by re-implementing the hierarchical LCC assignment scheme in iVia using Logistic Regression classifiers instead of Support Vector Machines, and using words as features instead of LCSH Heads. Because Logistic Regression is probabilistic, fewer component classifiers are required, resulting in a faster, simpler training and classification processes.

### 4.7.3 Media Type Assignment

iVia describes the format of each Internet resource with a single Media Type (also known as MIME type) value like *text/html* or *application/pdf* [14]. If an HTTP header is available for a resource, iVia attempts to assign a Media Type based on the *Content-type* field. In the absence of a header, the Media Type is assigned by using the *libmagic* library (which is the basis of the common Unix *file* command) to analyze the document. *libmagic* determines the Media Type of a using a database of rules that map the unique low-level features of numerous different types to their Media Type. For example, the rule

```
0 string %PDF- application/pdf
```

states that if position 0 in a file contains the string *%PDF-*, then the Media Type of the file is *application/pdf*. On the author's Linux workstation, there are 308 rules, which identify 183 different file types.

## 5. DISCUSSION

The evaluations described above are development tools, and form an invaluable guide to whether changes made to assignment algorithms have positive or negative effects. However, they have some limitations. Like all automatic evaluations, these statistics do not fully account for near-misses. This drawback can usually be overlooked: if a change to an assignment process results in the measures of quality increasing, we can assume it is a positive change. However, there is an ongoing risk that we will design an algorithm to maximize the quality measures, not to produce good metadata. We are therefore supplementing our automatic evaluations with a set of human evaluations, to be performed later this year.

Occasionally, the INFOMINE editors decide to overrule the recommendations made by the automatic metadata assigner. For example, *PhraseRate* is designed to assign no single-word phrases, even though 49% of the expert-assigned keyphrases in INFOMINE are a single word, and this restriction limits recall to 51%. However, the editors' independent evaluations show that this restriction improves output because single-words, while appearing to match the quality of longer phrases, can contain very visible and distracting mistakes. Several similar decisions have

been made, including the special rule for government publications in INFOMINE Category Assignment (Section 4.6.2), and the order of Title metadata sources (Section 4.1.3).

Our ongoing work includes new assignment tools and evaluation metrics. We also plan to calculate confidence intervals for each statistic, though this will require a random sampling procedure instead of the current most-recently-modified sampling.

## 6. CONCLUSION

Automatic metadata assignment is becoming more and more vital, as virtual libraries, metadata repositories and other digital libraries attempt to impose order on the Internet. This paper has described a set of tools for assigning many common metadata fields to Internet resources. Some of the tools are almost trivially simple, while others are very complex. In combination, they are able to supply an abundance of descriptive metadata which can be used in a variety of situations for a range of purposes.

As we have developed the automatic metadata assignment tools, we have also developed an automatic metadata evaluation tool to measure and guide their progress. Different metadata fields contain different types of data, and should be measured by different statistics. We have chosen a set of measures appropriate to each field, and evaluated the tools we use, and others we have trialled, using these measures. The metadata evaluation tools have proved their value on numerous occasions, though we stress that they are imperfect surrogates for formal human evaluations.

## 7. REFERENCES

- [1] Barker, K. and Cornacchia, N. Using Noun Phrase Heads to Extract Document Keyphrases. In *Proc. Thirteenth Canadian Conference on Artificial Intelligence (LNAI 1822)*. Montréal, Canada, 2000, 40-52.
- [2] Cavnar, W. B. and J. M. Trenkle, N-Gram-Based Text Categorization. In *Proc. Third Annual Symposium on Document Analysis and Information Retrieval*. UNLV Publications/Reprographics, Las Vegas, NV, 1994, 161-175.
- [3] Chakrabarti, S., Roy, S. and Soundalgekar, M. V. Fast and Accurate Text Classification via Multiple Linear Discriminant Projections. In *Proc. VLDB*. 2002, 658-669
- [4] Chan, L. M. Inter-indexer consistency in subject cataloging. *Information Technology and Libraries*, 8, 4. 1989, 349-358.
- [5] Dolin, R. A. *Pharos: A Scalable Distributed Architecture for Locating Heterogeneous Information Sources*. Ph.D. Thesis, University of California, Santa Barbara. 1998.
- [6] Dublin Core Metadata Initiative *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. 1995-2005. <http://dublincore.org/documents/dces/>
- [7] Frank E., Paynter G.W., Witten I.H., Gutwin C. and Nevill-Manning C.G. Domain-specific keyphrase extraction. In *Proc. IJCAI*, Morgan Kaufmann, 1999, 668-673.
- [8] Frank, E., Hall, M., and Pfahringer B. Locally Weighted Naive Bayes. In *Proc. Conf. Uncertainty in Artificial Intelligence (UAI 2003)*. 2003, 249-256.
- [9] Frank E. and Paynter, G. Predicting Library of Congress Classifications From Library of Congress Subject Headings. *JASIST*, 55, 3. 2004, 214-227.
- [10] Godby, C. J. & Stuler, J. The library of congress classification as a knowledge base for automatic subject categorization. In *Subject Retrieval in a Network Environment: Papers Presented at an IFLA Satellite Meeting Sponsored by the IFLA Section on Classification and Indexing and IFLA Section of Information Technology*, (Dublin, OH.). OCLC, 2001, 14-16.
- [11] Guy, M. Powell, A. and Day, A. Improving the Quality of Metadata in Eprint Archives. *Ariadne* 38, January 2004. <http://www.ariadne.ac.uk/issue38/guy/>
- [12] Han, H., Giles, C. L., Manavoglu, E., Zha, H. Zhang, Z. and Fox, E. Automatic Document Metadata Extraction using Support Vector Machines. In *Proc. JCDL*. 2003, 37-48.
- [13] Humphreys J. B. K. *PhraseRate: An HTML Keyphrase Extractor*. Technical report, University of California, Riverside. June 2002. [http://infomine.ucr.edu/projects/Keith\\_Humphrey/PhraseRate/phraserate.pdf](http://infomine.ucr.edu/projects/Keith_Humphrey/PhraseRate/phraserate.pdf)
- [14] Internet Assigned Numbers Authority. *MIME Media Types*. <http://www.iana.org/assignments/media-types/>
- [15] Jones, S., Lundy, S. and Paynter G. W. Interactive document summarisation using automatically extracted keyphrases. In *Proc. Hawai'i International Conference on System Sciences: Digital Documents: Understanding and Communication Track*. IEEE-CS, 2002, 101-109.
- [16] Jones, S. and Paynter G. W. Human evaluation of Kea, an automatic keyphrasing system. In *Proc. JCDL*. ACM Press, 2001, 148-156.
- [17] Jones S. and Paynter G. W. Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. *JASIST*, 53, 8. 2002, 653-657.
- [18] Jones S. and Paynter G. W. An evaluation of document keyphrase sets. *Journal of Digital Information*, 4, 1. 2003. <http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Jones/>
- [19] Kedzierski, A. *Artur's Auto Annotator*. Masters Thesis, Department of Computer Science, University of California, Riverside. 2002.
- [20] Komarek, P. *Logistic Regression for Data Mining and High-Dimensional Classification*. Doctoral Thesis, Department of Mathematical Sciences, Carnegie Mellon University. 2004.
- [21] Larson, R. R. Experiments in automatic library of congress classification. *JASIS*, 43, 2. 1992, 130-148.
- [22] Library of Congress. *SuperLCCS: Library of Congress Classification Schedules combined with additions and changes*. Gale Research Inc. 1986-2001.
- [23] Library of Congress Subject Cataloging Division. *Library of Congress Subject Headings (24 Ed.)*. Library of Congress. 2001.
- [24] Lin, C. and Hovy, E. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proc. Human Technology Conference*. Edmonton, Canada, 2003.
- [25] Lovins J. B. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11. 1968, 22-31.
- [26] McCallum, A., Rosenfeld, R., Mitchell, T. and Ng, A. Improving text classification by shrinkage in a hierarchy of classes. In *Proc. Fifteenth International Conference on Machine Learning (ICML-98)*. 1998, 359-367.
- [27] Mitchell S., Mooney M., Mason J., Paynter G. W., Ruschinski J., Kedzierski A., Humphreys K. iVia Open Source Virtual Library System. *D-Lib Magazine* 9, 1. January 2003. <http://www.dlib.org/dlib/january03/mitchell/01mitchell.html>
- [28] Ruiz, M. E. and Srinivasan, P. Hierarchical neural networks for text categorization. In *Proc. SIGIR*. 1999, 281-282.
- [29] Turney, P. D. Learning Algorithms for Keyphrase Extraction. *Information Retrieval* 2, 4. 2000, 303-336.
- [30] Turney, P. D. Coherent Keyphrase Extraction via Web Mining. In *Proc. IJCAI*. 2003, 434-442.
- [31] Witten, I.H. and Bainbridge, D. *How to Build a Digital Library*. Morgan Kaufmann, San Francisco, CA. 2003.
- [32] Witten, I. H. and Frank, E. *Data Mining*. Morgan Kaufmann, San Francisco, CA. 2000.
- [33] Yilmazel, O. Finneran, C. M., Liddy E. D. Metaextract: an NLP system to automatically assign metadata. In *Proc. JCDL*. 2004, 241-242.