

# Automatic Labelling of References for Internet Information Systems

A. Geyer-Schulz, M. Hahsler

Abteilung für Angewandte Informatik insbesondere Betriebsinformatik,  
Wirtschaftsuniversität Wien, A-1090 Vienna, Austria

**Abstract:** Today users of Internet information services like e.g. Yahoo! or AltaVista often experience high search costs. One important reason for this is the necessity to browse long reference lists manually, because of the well-known problems of relevance ranking. A possible remedy is to complement the references with automatically generated labels which provide valuable information about the referenced information source. Presenting suitably labelled lists of references to users aims at improving the clarity and thus comprehensibility of the information offered and at reducing the search cost. In the following we survey several dimensions for labelling (time, frequency of usage, region, language, subject, industry, and preferences) and the corresponding classification problems. To solve these problems automatically we sketch for each problem a pragmatic mix of machine learning methods and report selected results.

## 1 Introduction

Internet Information Systems (IIS) such as Internet search-engines used today provoke high search costs. The user has to spend a lot of time checking enormous lists produced by the systems to find relevant items. For example, for the query of the phrase +Marketing +Course AltaVista (Compaq (1999)) found more than 800.000 hits without use of meta information (October 1999). It is impossible for the user to exhaustively search such vast amounts of information sources manually.

To cope with this problem, search-engines use relevance ranking to order their output after performing a boolean search. Most implementations are based on the vector space model of Salton, McGill (1983) to compute a similarity measure between the query and the documents found. But since queries for IIS tend to be very short and often consist of only one or two terms, the similarity measure degenerates to a simple frequency measure and, therefore, provides only limited help to the user. Some alternative approaches to classical ranking algorithms are:

1. Use other measures to sort the output (e.g. cross citation index, usage frequency, user recommendations ...).
2. Show additional information to enable the user to browse the list faster.

In computer vision labelling techniques and algorithms have a long tradition for constraint propagation (e.g. Cohen, Feigenbaum (Eds.) (1982)) or for labelling image parts (e.g. Leung, Yang (1995)). However, in this paper we understand

labelling as a visualization technique for Internet resources. Section 2 describes the labelling process. Next, we discuss several dimensions of labelling like time, region, language, rating and resource type. In section 4 we present the results of testing user acceptance of labelling and in section 5 we summarize benefits and drawbacks of labelling.

## 2 Labelling

A label is the visualization of a special property of an information object (e.g. a link returned by the IIS) that represents additional meta information about it. Meta information is a description of the information object that contains certain aspects of its resource type, subject, language, relation to other objects . . . (Dublin Core (1999), W3C (1999)).

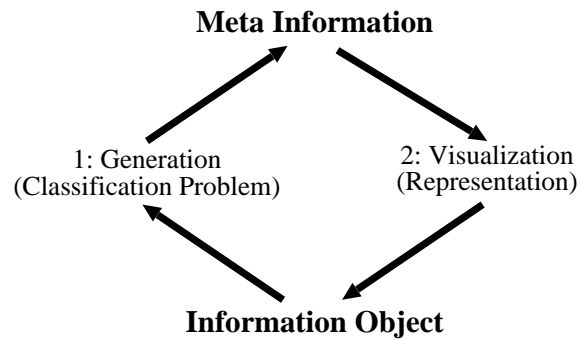


Figure 1: Labelling Process

For each label the following steps are necessary (Figure 1):

1. Generate meta information. For each type of meta information this represents a classification problem: In a first step, the classes most helpful for the user must be identified, then the information objects must be assigned to the appropriate class. This is done either at the time an object is added to the IIS, at regular intervals or, if necessary, on demand. Motivation for automating this tasks are in the considerable maintenance cost for large collections of information objects.
2. Visualize meta information for the user. Colored markers are appended to the items returned by the search-engine to allow the user to check the list faster. A further improvement can be achieved by e.g. offering dynamically refined lists that only contain items of a certain type of meta information.

An important problem is to understand how users exploit what kind of meta information to reduce their search time in retrieval tasks and which visual representation of meta information is preferred or best understood by users.

## 3 Dimensions of Labelling

A dimension of labelling is a set of labels that are available to inform the user about a specific property of an information object. There are general dimensions representing properties useful for all kinds of IIS (e.g. the time dimension), but there also exist dimensions that are dependent on the type and structure of the information objects in a particular IIS (e.g. the dimension industry used for a collection of companies) and on the specific context of the IIS. In the following we describe some dimensions and sketch pragmatic approaches for solving the underlying classification problems automatically.

### 3.1 Dimension Time

The time dimension informs the user about the state of an information object, or better information product, in its life-cycle. Like other products, information products are produced, they have a usable lifetime, and finally they become obsolete. During their lifetime information products can be re-launched to extend their usable life. To represent the possible states of an information object, we use the following categories:

**New** - The information object was recently added to the IIS.

*Use Case:* Users want to stay up-to-date with the system and an easy way to identify new objects.

*Solution:* Store the submission date and use the revisit frequency of the average user to determine the time span during which the label is shown.

**Revised** - The content has changed (product re-launch).

*Use Case:* Users want to stay up-to-date with a particular information object and track changes of its content.

*Solution:* Use a robot to check the information object frequently for changes. Determine the time span to show the label as in the previous case.

*Problem:* In a hyper-text environment the delimitation of one information object (consisting of several, maybe dispersed web pages) from others is a non-trivial problem.

**Obsolete** - The information object is no longer maintained.

*Use Case:* Users do not want to find references to non-existing objects.

*Solution:* Use a robot to check the information object frequently (using sampling techniques) and delete references to objects no longer available.

*Problem:* Unreachability may be caused by temporary problems. Technological problems are, for example, down-times due to problems with infrastructure or maintenance. From an organizational point of view, the whole object could be moved to another location, which means that the reference to the object has to be corrected.

## 3.2 Dimension Region and Language

By using a geographical dimension the user can easily identify items with information about a specific region. This is, for example, essential to provide the user with a regional overview of an industry. Another vital property of an information object is its language. It shows whether the information is understandable and thus usable for the user. If the language is identified, the IIS can also offer automatic translation services to make even information in a foreign language accessible. The labels used are:

**Region** - Is the information object specific for a certain region?

*Use Case:* A user searches for companies operating in a particular country or economic region.

*Solution:* Use the country code in the URL as specified in ISO 3166 (1997) or the contact information provided by the HTML <address>-tag.

*Problem:* This is only a very rough solution since some top-level domains (like .com, .edu ...) do not carry country information and the <address>-tag can only rarely be found in today's web pages. And even if it is found, it is hard to analyze. Other unresolved issues are: Disparity between geographical and political regions (e.g. Hawaii), level of detail of address representation (e.g. U.S.A. - California - San Diego).

**Language** - Is the information understandable for the user?

*Use Case:* A user searches for newspapers in a particular language to improve his language skills.

*Solution:* Use statistical or probabilistic methods for language identification (e.g. n-gram transition probabilities used in Dunning (1994)) to verify the information in the HTML 'lang'-attribute or the HTTP header field 'content-language'.

*Problem:* Authors of web pages have a specific way of language use. There is a tendency towards incomplete sentences, the usage of Internet words like "Homepage" or "Feedback". Many pages are bilingual. These factors complicate language identification.

## 3.3 Dimension Rating

Dimensions like time, region and language represent properties directly derived from information objects. Rating, in contrast, addresses aspects that are assigned to the object by its users and can not directly be derived. Usage, for example, changes over time while the respective information object stays the same. Thus the result of rating is not directly a property of the object, but it rather shows how frequent, in the case of usage, the user chooses a specific object. Other labels rate the quality of infrastructure, which means the probability that the information object can be reached or the average time needed to get the object. The last rating label presented here visualizes explicit recommendations by users.

**Usage** - Which information is frequently selected by users?

*Use Case:* Users want to know which information source is preferred by

other users.

*Solution:* Observe and compare the usage frequency of information objects with similar content. Do an ABC-analysis for the objects and label all A-objects with “Hot” or visualize utilization with a bar graph.

*Problem:* Known problems with observation on the Internet are: It is hard to distinguish between usage by real users making a choice and access by robots that only follow their internal rules and do not care about content. Also the impact of caching (by Web-browsers and proxies) on the usage measure is hard to estimate. Furthermore a simple usage measure only works with homogeneous user groups. If we can identify several user groups, usage frequency has to be evaluated separately for each group. This requires a personalized interface.

**Infrastructure** - Is the source of an information object technically reliable?

*Use Case:* Users are interested in how fast they will get the information object and the probability to get it at all.

*Solution:* Use a robot to check the information object frequently and classify response time and frequency of failure. Visualize reliability for example like Fast (1999) with a number of stars.

**Recommendations** - Which objects are suitable and suggested for special target groups (e.g. novice - intermediate - expert)?

*Use Case:* Novices in a certain subject want to know which information sources other, more experienced users or experts recommend.

*Solution:* Provide forms for on-line rating.

*Problem:* Explicit collaborative rating by users can hardly be automated and requires active participation of users. The group dynamic nature of this social process can be supported by functionalities known from group-ware applications. Trust in the rating is a key factor here. A recommendation by a well-known expert may be of more importance than the rating by a group of unknown users.

### 3.4 Dimensions Without Simple Classification Rules

All labels presented above are based on relatively simple classification problems and quite good results can be obtained with simple rules. But to determine the type of an object (e.g. the object can represent a lecture, a journal, a company home page, or just a private collection of links) there exist no simple rules. For such classification problems we need a flexible way to learn the required rules and to adapt them, if necessary.

A machine-learning approach which gradually reduces human involvement for this type of classification problem consists of the three steps shown in figure 2. In the first stage new information objects are assigned manually to predefined categories. After a sample of sufficient size is entered into the system, cluster centers based on the vector space model of IR (Van-Rijsbergen (1979)) are computed for each category. These clusters are used in stage 2 to categorize new objects automatically. The result of this preliminary automatic classification is checked manually and rules are incrementally updated with the feedback (supervised or

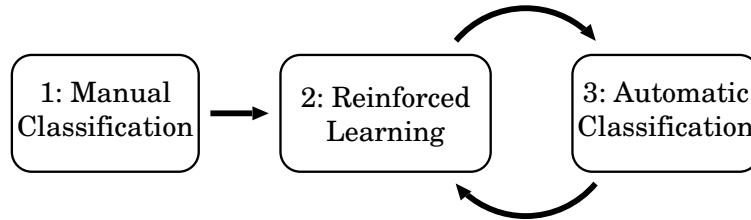


Figure 2: From manual to automatic classification

reinforced learning). See Sutton (1992). After the error rate of the automatic classification process falls below a certain threshold value, the system proceeds to stage 3. In this stage only a sample of all automatically processed objects is checked manually. If the sample error rate grows above a certain threshold, the system switches back into the reinforced learning mode.

## 4 Experimental Results

As a test case for labelling we use the Internet Information System developed for the *Living Lectures – Virtual University Project* (VU (1999)) at the *Vienna University of Economics and Business Administration*. The aim of this IIS is to provide access to Internet-based information to support learning, teaching and research at the university. By February 1999 the IIS contained references to more than 4.000 information objects dispersed all over the Internet.

We questioned 40 persons (students, faculty and administrative staff) about the IIS. 67% of the test participants knew the system and usage frequency ranged almost uniformly from never to several times a week. So we had unexperienced as well as heavy users in the sample. The questionnaire confronted the participants with the hypothetical task to search for an appropriate HTML-tutorial for creating or expanding their own home page. We showed the test participants where to find a list of HTML-tutorials (a similar list is shown in figure 3) and asked them to consider the items on the list. Then we asked them, if they noticed the colored markings. Not surprisingly, 92% noticed. The next question was whether the concept behind a label is understandable. 67% of the test participants linked the "Revised"-label intuitively with the right concept of marking recently changed information objects. Only 20% of all participants felt that the used labels were very helpful, but cross tabulation shows that the perceived utility of labelling increases with usage frequency. Heavy users found the labels much more helpful than participants who used the IIS never or very seldom. This suggests that users have to get used to the labels and learn their meaning to benefit from the additional information they convey.

Entries found for *CATEGORY:='Dictionaries': 25*

Name	Source	Category	Description
<u>LEO: German &lt;--&gt; English</u> <i>(no frame)</i> English ■ (48%)	Fuhrman, B., Bodzin, A., Jung, A. Maurer, H.	Dictionaries	<a href="#">Description</a>
<u>A Web of On-Line Dictionaries</u> <i>(no frame)</i> English ■ (14%) META	Robert Beard	Dictionaries	<a href="#">Description</a>
<u>Eurodictionary</u> <i>(no frame)</i> English ■ (9%)	©ECSC-EC-EAEC, Brussels-Luxembourg, 1996, 1997	Dictionaries	<a href="#">Description</a>
<u>Lexika (Universität Erlangen)</u> <i>(no frame)</i> German ■ (8%) META	Institut für Germanistik	Dictionaries	<a href="#">Description</a>
<u>Diccionarios digitales</u> <i>(no frame)</i> Spanish ■ (4%) REVISED	LA PÁGINA DEL IDIOMA ESPAÑOL	Dictionaries	<a href="#">Description</a>
<u>Glossar Internet und bibliothekarische</u>			

Figure 3: Sample from the Living Lectures – Virtual University Project

## 5 Benefits and Drawbacks of Labelling

The main benefit of labelling is the additional information available to the user to make his choice. The colored labels are easy to recognize and give a general impression of what kind of information object the user will find behind the reference and in which condition it currently is. Labelling permits additional functionality, e.g. dynamically generated “What’s New” or “What’s Revised” lists. These features drastically reduce the time users need to track information in the IIS.

Another benefit of labelling is the possibility to increase precision and recall of the IIS. Filtering unwanted results increases precision, categories as search terms increase recall. Furthermore, categories can be used to make the IIS browsable by selecting intersections between several categories.

The experimental results reveal that a single colored marking (a label) is often used to convey complex concepts (e.g. usage). Such markings can easily be misinterpreted. This leads to annoyed users.

A similar problem is sensory overload, or the question how much and which labels should be presented to the user. Important features which influence sensory

overload are the shapes, the number of colors and the arrangement of labels. A possible solution is to prepare several labels and hide unnecessary labels depending on the context. E.g. there is no need to show a “New”-label for a list created from the query “What’s New”.

Labels as well as all additional information added to a list tend to manipulate the choice process of users. Since all labels are generated automatically based on generally known rules, manipulation is kept at a minimum. For the list of online-dictionaries in figure 3 we use usage (represented by the bar and the percentage in brackets) to rank the output. So frequently used dictionaries migrate slowly to the top of the list. There they receive even more attention because of their prominent position. However, when properly done, this is just a process to promote dictionaries that are useful to users of the IIS.

## References

COHEN, P.R. and FEIGENBAUM, E.A. (Eds.) (1982): *The Handbook of Artificial Intelligence*, Vol. 3, Chapter XIII: Vision, William Kaufmann, Los Altos, California.

COMPAQ (1999): AltaVista. Online: <http://www.altavista.com/>.

DUBLIN CORE (1999): Dublin Core Metadata Initiative.  
Online: <http://purl.oclc.org/dc/>.

DUNNING, T. (1994): *Statistical Identification of Language*. Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University.

FAST (1999): Fast MP3 Search. Online: <http://mp3.lycos.com/>.

ISO 3166 (1997): *Codes for the Representation of Names of Countries and their Subdivisions*. International Organization for Standardization.

LEUNG, M.K. and YANG, Y.-H. (1995): First Sight: A Human Body Outline Labeling System, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4), 359–377.

SALTON, G. and MCGILL, M.J. (1983): *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

SUTTON, R. S. (1992): Introduction: The Challenge of Reinforcement Learning. *Machine Learning*, 8(3/4), 225–227.

VAN-RIJSBERGEN, C.J. (1979). *Information Retrieval*. Butterworths, London. 2nd edition.

VU (1999): Living Lectures – Virtual University Project.  
Online: <http://vu.wu-wien.ac.at/>.

W3C (1999): *Resource Description Framework (RDF) Model and Syntax Specification*. W3C Proposed Recommendation.  
Online: <http://www.w3.org/TR/PR-rdf-syntax/>.