Retrieval of Italian legal literature: a case of semantic search using legal vocabulary

E. Francesconi, G. Peruginelli
ITTIG – Institute of Theory and Techniques of Legal Information
Italian National Research Council,
Florence, Italy
e-mail: {francesconi, peruginelli}@ittig.cnr.it

Abstract:

Retrieval of legal information and in particular of legal literature is examined in conjunction with the creation of the Portal to Italian legal doctrine developed by the Institute of Theory and Techniques of Legal Information (ITTIG) of the National Research Council of Italy. Subject searching is a major requirement for Italian legal literature users and a solution is described for the retrieval of legal literature's resources. Such solution is based on the exploitation of Dublin Core metadata for both data coming from structured repositories and for web documents, as well as the use of a controlled vocabulary list prepared for accessing indexed articles of the DoGi – Dottrina Giuridica database. Technical specifications are illustrated as well as advantages and limitations of such solution.

Keywords:

Legal literature, precision and recall in legal information retrieval, controlled vocabulary.

1. Introduction

Legal information has specific features due to its nature, its different utilisation purposes and the intrinsic need for integration of its components, represented by legislation, cases and literature. Access to legal literature[1] in particular is a primary requirement: it responds to the demand for understanding and interpretation of statutes and cases, an objective that law scholars and professionals greatly contribute to.

The difficulty of delimiting the scope of legal literature's sources makes its access problematic in comparison to other legal information sources. Retrieving legal doctrine necessitates a long and cumbersome research activity across multiple sources as there is no one single information provider that legal researchers can effectively gain access to.

For this reason the Institute of Theory and Techniques of Legal Information of the National Research Council (ITTIG-CNR) puts efforts in developing a system to ensure a unified point of access to legal literature: the Portal to legal literature (1) (2).

The architecture of the Portal deals both with data coming from structured data repositories and with web documents. The aim of the Portal is to integrate these two different data sources in a unique view using Dublin Core metadata.

Data of structured repositories are essentially bibliographic metadata which are harvested, at service provider level, using the OAI-PMH approach and mapped from the native format into Dublin Core metadata scheme (1). On such data the Portal retrieval system allows users to search by subject metadata following a semantic approach, since the semantics of

[1] Legal literature in Civil law systems, consists in legal intellectual output which is found in monographs, journal articles, manuals, grey literature, proceedings, etc.

https://doi.org/10.23106/dcmi.952108070

these data is clearly defined.

On the other hand, web documents usually lack specific metadata as well as reliable or uniform HTML meta-tags, which can help the qualification of documents. Such kind of data are essentially represented by natural language text, whose structure and semantics are not univocally defined.

In order to integrate different data sources in a unique view, web documents have to be provided with a DC metadata description as well. This is what has been implemented so far, in two different phases, for the Portal to legal literature.

The current phase of the project is dedicated to the enhancement of the existing solutions oriented to semantic searching, as well as to the implementation of specific facilities to support legal users in semantic querying the Portal, trying to guarantee both precision of retrieval and recall efficacy.

2. Legal users' information needs

Users' information needs can be defined as a gap between what we know and what we want to know that motivates a search (3): this results in the formulation of a query. Users' information needs are recognised as an essential factor in the information seeking process (4).

Expressing information needs by users is rarely a straightforward process due to lack of clear identification, in formulating queries, of the exact terms expressing legal concepts (5).

Users of legal literature share the characteristics, attitudes and needs of other users in seeking information, but they have some peculiarities due to the sophisticated nature of legal information. In particular Italian legal users are mostly interested in subject access facilities. Legal concepts are generally expressed both in natural and technical language and for this reason they must be provided with adequate semantic tools helping them to contextualize information, and enhance searching performances.

Legal users should be provided with semantic tools (controlled vocabulary and automated indexing tools) enabling adjustment and reformulation of their information needs, helping them to better identify their requirements and consequently expanding their queries.

3. Meeting the information needs of the Portal's users

As discussed in Section 2, to match legal user information needs at its best, the Portal has to provide high quality services able to guarantee both precision and recall.

In particular, to guarantee retrieval precision the Portal aims at enriching documents with high quality metadata, so that retrieval is more focused and able to better match the semantics of the query.

Moreover, to guarantee recall in retrieval the Portal aims at matching, with the query, the related information needs. To obtain this, the query has to be formulated in a way that express at its best the semantics of such needs. For this purpose users are to be offered facilities to construct a query, browsing a hierarchy of legal categories, as well as to expand it with broader or narrower terms. Such expansion of a user's original query can reliably retrieve relevant documents which did not match the query as originally formulated.

For example, let us consider a query searching for documents of the legal category "criminal law" where the term "recklessness", correlated to the chosen legal category is included. If relevant documents for the query do not contain the term "recklessness" a 'no hits' response is given back. The expansion of the query with broader or narrower terms, as found in the legal controlled vocabulary, related to the chosen legal category can lead to retrieve relevant documents.

This service improves the expressiveness and completeness of the query, so to match at its best user information needs, as well as to achieve retrieval recall.

In order to provide semantic search services meeting the goals of both precision and recall in retrieval, the Portal architecture has been designed (Fig. 1):

on data side

to harvest and map metadata of structured repositories into DC metadata scheme using the OAI-PMH approach;

to provide Web documents with a semantic description according to the DC metadata scheme using an automatic metadata generator (1);

thus providing a uniform view on data;

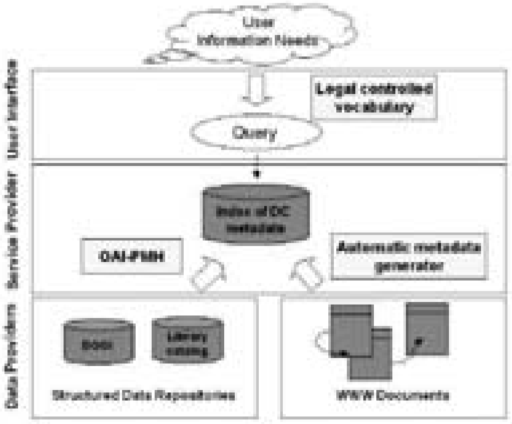


Fig. 1 The Portal services oriented to semantic search.

https://doi.org/10.23106/dcmi.952108070

on user side

to provide facilities helping users in formulating a query effectively expressing the semantics of their information needs, by using a high quality legal controlled vocabulary.

In Section 4 and 5 the needs of semantic searching in the legal domain, along with the tools used to improve, respectively, precision and recall in retrieval, will be thoroughly discussed. Section 6 introduces the technical solutions adopted for the Portal, deeply discussed in Section 7 and 8.

4. The semantics on data to enhance precision in retrieval

In order that legal literature can be effectively accessed, essential requirements are represented by quality of indexing and adequate retrieval facilities provided to users.

A survey of the use made of the major Italian legal bibliographic indexing service, the DoGi database[2] has been conducted, to find out users' behaviour in seeking information and their purpose for searching. The aim was to identify users' profile in order to plan services which better meet their needs. The major finding which has emerged is the paramount importance of semantic indexing of legal literature. For successful and high quality search and retrieval, legal literature documents have to be analysed in a way that the indexing language is consistent with that used by legislators and judges, while exploiting appropriate tools such as controlled vocabularies and authority lists.

In an electronic environment the requirement of quality indexing is very difficult to achieve. A great amount of legal literature resources available on the web are not indexed and when subject metadata are provided, the difference between diverse legal subject and classification systems is a serious obstacle to successful retrieval.

Moreover very few specialised areas of law are taken into account when a categorization is available. Some examples are: a) the area of criminology is quite often simply indexed as criminal law and not as a self-contained area of law; b) the area of environmental law, despite its relevance, is most of the times confined in the broader area of administrative law.

All this implies a major difficulty for users in retrieving information, who also miss the opportunity to have a clear and complete view of the various legal areas and of the overall structure of law of a particular legal system.

Such conceptual confusion hinders performing refined searchers according to expert users' needs and skills, while less experienced users are presented with a non systematic and eventually wrong view of areas of law.

Facilities for searching within the various areas of law by subdivisions or by specific concepts of a particular legal system are quite uncommon. Some specialised sites focusing on a single area of law do provide specific sections showing resources on a given concept, but these resources are not individually searchable as they are not specifically indexed, but simply included in predefined sections.

Therefore the importance of human intermediation in filtering networked legal literature's resources information is a crucial factor, demonstrated by the widely recognized importance of one of the evaluation criteria of Web sites: the need for metadata to facilitate retrieval and effective use of legal literature.

However the creation of adequate tools such as controlled vocabularies is a time-consuming and expensive activity so that attempts have been made to automatically index legal literature's electronic resources. In this direction the Portal has adopted an approach for automatic indexing of web legal literature available on the Internet (1).

In order to supply web documents with metadata, an automatic metadata generator module based on a machine learning approach has been developed for the Portal. This module aims at supporting the intellectual activity of a service provider in its efforts of organizing electronic legal literature's resources. In particular a "*dc:subject*" metadata generator has been constructed (1). Considering the importance of the classifier module, in this phase of the project, attempts have been made to enhance the accuracy of such classifier, so that semantic search services can be based on reliable metadata indexing (see Section 7).

5. The DoGi classification system to improve recall in retrieval

It is our opinion that through the exploitation of a controlled legal vocabulary users can satisfactorily retrieve relevant materials compensating the lack of match between queries and terms contained in the full text of documents.

At the moment users accessing the Portal can enter a search term as well as choose a given area of law to contextualise their search. Such terms are matched against the full text of documents. The absence of such term in the text inevitably leads to no results. In

[2] The DoGi database (http://nir.ittig.cnr.it/dogiswish/Index.htm), produced and distributed by the Institute of Theory and Techniques of Legal Information, which forms part o the National Research Council (ITTIG-CNR) is, in the Italian legal landscape, one of the most precious sources for legal literature research. It is a database created in 1970, offering abstracts of articles published in the most important legal periodicals (more than 250).

https://doi.org/10.23106/dcmi.952108070

order to overcome this limitation, a solution has been devised to help users navigating through resources they could hardly retrieve.

Such solution is based on the exploitation of a legal controlled vocabulary. At present retrieval services of legal literature in Italy is mainly provided by libraries, private and public information centres and by a small number of publishers. For these information providers setting up joint and effective legal information services is quite a difficult task. They use legal classification schemes which vary in scope and methodology. Multidisciplinary classification systems such as the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) and the Library of Congress Classification (LCC) are mostly used for indexing legal material. Legal classification schemes designed for use by legal communities are far less popular.

The use of multidisciplinary classification systems in the area of law may seriously limit the retrieval of relevant information. For example the Dewey classification, in its effort to arrange knowledge into predefined classes, sometimes misses to identify specific concepts. Despite the numerous revisions and the attempts to internationalise the scheme, law is still much oriented to the common law system and problems arise when trying to fit some legal concepts which are peculiar to the civil law system and in particular to Italian law, into the 34X notation.

The DoGi classification system is able to overcome some of the problems mentioned above.

The mostly used Italian legal classification system is in fact the one adopted for indexing DoGi documents. As already mentioned the goal of DoGi database is to provide law scholars and professionals with exhaustive and updated legal information as found in Italian peer-reviewed articles on law.

The indexing language is a controlled one and is based on the areas of law as structured in the Italian law faculty scheme. Such classification is a valid tool not only for retrieving legal literature items in the DoGi database, but also for an in- depth understanding of the structure of Italian law. There are 24 areas of law considered, each designated by a code.

The classification scheme is hierarchically structured (up to three levels) and is composed by alphanumeric codes expressing specific concepts. Codes are associated with descriptors (6600 at the moment). An authority list of descriptors is maintained and updated on the basis of indexers' suggestions, as well as of statistic analysis of searches made by users. The classification scheme has been conceived as a dynamic instrument which is periodically reviewed and new codes are established reflecting additional topics dealt with in the literature. This is an example of the structure of the area of European law.

UNEUR.0. European Union law
UNEUR.1. European Union
UNEUR.1.0. Institutions and other bodies of European Union
– Committee of the Regions
– Council of the European Union
– Court of Auditors
– Court of Justice
– European agencies
– European Central Bank
– European Commission
– European Data Protection Supervisor
– European Economic and Social Committee
– European Ombudsman
– European Parliament
UNEUR.2. European Community (First Pillar)
UNEUR.3. Common Foreign and Security Policy – CFSP (Second Pillar)
UNEUR.4. Police and Judicial Co-operation in Criminal Matters (Third Pillar)

The potential of such controlled vocabulary, considered in this context as an independent self-contained semantic tool, consists in the fact that it is a way to introduce an interpretative layer of semantics between the term entered by the user and that present in the controlled vocabulary itself. It adds to the research process by suggesting related terms, thus expanding the possibility of locating concepts. This is made possible as in the controlled vocabulary entries are presented under main headings along with associated concepts.

The solution under development follows an approach in which search terms that are not contained in the full text of documents are matched against the controlled vocabulary list, thus helping users learn the structure of legal concepts and the related terminology. This is made possible by a query interpretation module based on approximation, pointing to narrower or broader terms that are used to automatically launch a search in the full text documents, while respecting the specific area of law originally requested by users.

6. Solution implemented for semantic search

As discussed in Section 3, the solutions adopted for the Portal to obtain reliable semantic search services aiming to improve both precision and recall in retrieval, can be distinguished (Fig. 1) between those ones adopted on data side and those on user side.

As regards the data side, in the prototype of the Portal (1) solutions have been implemented to provide documents with a uniform semantic description according to the DC metadata scheme, by using the OAI approach for data coming from structured repositories and an automatic metadata generator for Web documents.

https://doi.org/10.23106/dcmi.952108070

As regards the automatic metadata generator, methodologies based on document properties (for example the document url for *dc:identifier*, the content of html tag <title> for *dc:title*, Web domain name for *dc:publisher* (see (1) for details)) and on machine learning approach for document classification (*dc:subject*) have been used (1).

In this phase of the project particular attention has been addressed to *dc:subject*. Different solutions for automatic document classification have been tested and in Section 7 a comparison between two different machine learning techniques (one of them also used for the first prototype of the Portal) are presented.

As regards the user side, a solution based on the use of a controlled legal vocabulary able to guide users in formulating queries well expressing their user information needs is discussed in Section 8.

7. Automatic legal Web document classification

The automatic Web document classifier implemented for the Portal mainly consists of a text categorization algorithm which takes as input the plain text of a Web document *d* and outputs its predicted type (or "class") *c* choosing from a set of candidate classes *C*. In order to perform such an operation, it relies on a machine learning algorithm which has been trained on a set of training documents *D* with known class, and thus learned a model able to make predictions on new unseen documents. A wide range of machine learning approaches have been applied to automated text categorization, and a vast literature on the subject exists (see (6) for a comprehensive review). Two correlated problems must be addressed in facing such a task: the choice of the document representation, that is how to turn the document into a format amenable for computation, and the choice of the particular learning algorithm to employ.

In Section 7.1 we present in details the different types of document representation that we have tested, while in Sections 7.2 we describe the two learning algorithms that were employed, *Naïve Bayes* (used in the first prototype of the Portal) and *Multiclass Support Vector Machines* (MSVM).

Finally, Section 7.3 reports an experimental comparison of the different methods and representations proposed.

7.1 Document representation

A number of alternatives are possible in order to represent a document in a format which can be managed by an automatic classifier. Two main problems have to be faced: the choice of the meaningful textual units, representing the atomic terms of the document, and the level of structure to be maintained when considering the combination of such terms. Concerning the second problem, the most common approach, which we followed in our implementation, is that of ignoring the sequential order of the terms within a given document, and representing it simply as an unordered bag of terms. Concerning the first problem, the simplest possibility is that of representing words as terms, but more complex approaches can be conceived. A number of authors (7) (8) have tried using phrases as terms, but their experiments did not produce significantly better effectiveness. We thus limited ourselves to individual words in our document representation. Nevertheless, a number of preprocessing operations have been tested on pure words in order to increase their statistical qualities and reduce the computational complexity of the problem:

- digit characters can be represented using a special character;
- non alphanumeric characters can be represented using a special character.

Other preprocessing operation as stemming or the use of word stoplists (stopwords), in this phase, have been considered.

Once basic terms have been defined, a vocabulary of terms *T* can be created from the set of training documents *D*, containing all the terms which occur at least once in the set. A single document *d* will be represented as a vector of weights , where the weight represents the amount of information which the term of the vocabulary carries out with respect to the semantics of *d*. We tried different types of weights, with increasing degree of complexity:

- a *binary* weight indicating the presence/absence of the term within the document;
- a *term-frequency* weight indicating the number of times the term occurs within the document, which should be a measure of its representativeness of the document content;
- a *tfidf* weight which indicates the degree of specificity of the term with respect to the document. Term Frequency Inverse Document Frequency is computed as

where D_w is the fraction of training documents containing at least once the term w. The rationale behind this measure is that term frequency is balanced by *inverse document frequency*, which penalizes terms occurring in many different documents as being less discriminative.

Moreover, statistics computed for extremely rare terms will be far less reliable, as already pointed out

https://doi.org/10.23106/dcmi.952108070

for phrases with respect to words, thus possibly leading to *overfitting* phenomena. In order to address such a problem, *feature selection* techniques can be applied to reduce the number of terms to be considered, thus actually restricting the vocabulary to be employed (6) (9). We tried two simple methods:

an unsupervised *min frequency* threshold over the number of times a term has been found in the entire training set, aiming at eliminating terms with poor statistics;

a supervised threshold over the *Information Gain* (10) of terms, which measures how much a term discriminates between documents belonging to different classes.

7.2 Classification Algorithms

Binary classification is a typical machine learning task, and a number of different approaches have been developed so far. Its extension to the multiclass case is straightforward for algorithms like decision trees (10), neural networks (11) or Bayesian classifiers (12), while algorithms like Support Vector Machines (13) require more complex extensions (see (14) for a review). In the prototype of the Portal (1) we employed a *Naïve Bayes* classifier, which proved quite effective for text categorization (16). In this work we have extended the data set to test the approach and to compare it with a multiclass extension of the *Support Vector Machines*.

The Naïve Bayes approach used in the prototype of the Portal has been described in (1). In this work the Support Vector Machines methodology to document classification is recalled according to a geometrical interpretation.

The first implementation of such a methodology, based on Kernel Methods (17) (18) for statistical learning (15) was that of Support Vector Machines (SVM) (13) (19) for binary classification tasks. Extensions to the multiclass classification case have been developed as either combinations of binary classifiers, or by directly implementing a multiclass version of the SVM learning algorithm (MSVM) (see (20), (14) for reviews and comparisons).

SVM for binary classification basically is a solution of an optimisation problem that attempts to find, among all the surfaces separating positive from negative examples of a training set, the surface by the widest possible margin (it is the middle hyperplane of the widest set of parallel decision surfaces separating positive from negative examples (Fig. 2)).

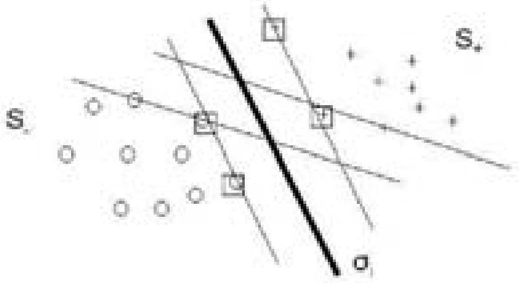


Fig. 2 Geometrical interpretation of the Support Vector Machine method for binary classification of linearly separable examples (S+ (S-) are the set of positive (negative) examples; i • is the decision surface, boxes indicates the support vectors).

Fig. 3 shows the geometric interpretation of the multiclass extension of the Support Vector Machine (MSVM) classifier, extended to most general case in which examples are not linearly separable. We can distinguish violations (training errors), occurring in case of non-linearly separable examples, and *support vectors*, which are the subset of training examples responsible for the learned decision function.

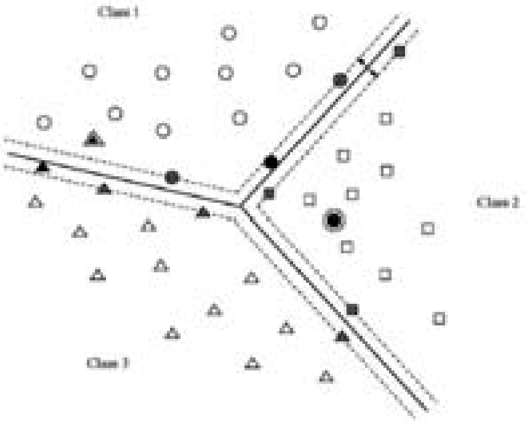


Fig. 3 Multiclass classification problem solved by MSVM. Solid lines represent separating hyperplanes, while dotted lines are hyperplanes which underline the confidence margin. The multiclass margin is the minimal distance between two dotted lines. Dark points are support vectors. Black points are also constraints violations and extra borders indicate violations which are also training errors.

For our experiments we used a direct implementation of the multiclass version of the SVM learning algorithm (MSVM) as developed independently by Vapnik (15) and Crammer and Singer (21).

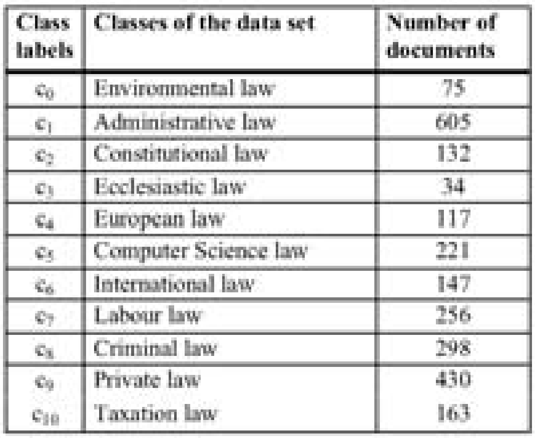
7.3 Experimental results

The two classification algorithms have been tested on a set of 2478 documents, belonging to the 11

https://doi.org/10.23106/dcmi.952108070

classes illustrated in Tab. 1. The documents have been selected from a set of Web sites of interest by a group of ITTIG legal experts. This data set has been used both to train and to test the classifiers.

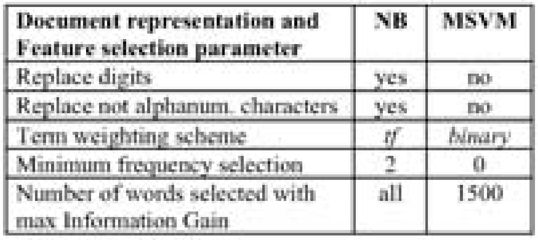
Tab. 1 Classes and number of documents for each class in the experiments



Class labels	Classes of the data set	Number of documents
c_0	Environmental law	75
c_1	Administrative law	605
c_2	Constitutional law	132
c_3	Ecclesiastic law	34
c_4	European law	117
c_5	Computer Science law	221
c_6	International law	147
c_7	Labour law	256
c_8	Criminal law	298
c_9	Private law	430
c_{10}	Taxation law	163

First of all a pre-processing step has been carried out aiming at removing html tags and javascript code within the documents. Then a number of combinations of the document representation and feature selection strategies have been tried. The parameters used for document representation and feature selection, which gave the best results for the two classification methods on our dataset are reported in Tab 2.

Tab. 2 Parameters for document representation and feature selection which produced the best results with Naïve Bayes (NB) and MSVM classifiers.



Document representation and Feature selection parameter	NB	MSVM
Replace digits	yes	no
Replace not alphanum. characters	yes	no
Term weighting scheme	*tf*	*binary*
Minimum frequency selection	2	0
Number of words selected with max Information Gain	all	1500

The first two rows represent possible preprocessing operations. The third row indicates the term weighting scheme employed. The two following rows are for feature selection strategies: the unsupervised minimum frequency and the number of terms to keep, after being ordered by Information Gain.

After having trained the Naïve Bayes classifier using the data set of Tab. 1, experiments have been carried out in order to validate such a learning procedure by evaluating the classification capability on the training set (*train accuracy*). The experiments produced the best results using parameters of Tab. 2 (column NB) for feature selection, obtaining a train accuracy of 82.5%.

Then, using the same data set a MSVM classifier[3] has been trained and tested. The best results have been produced using the parameters of Tab. 2 (column MSVM), obtaining a train accuracy of 85.1%.

A comparison of the results of the two classifiers are shown in Fig. 4.

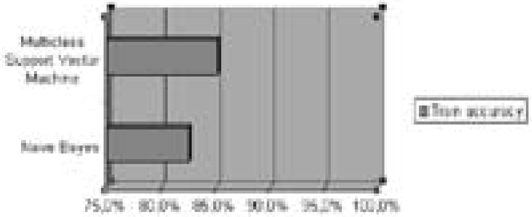
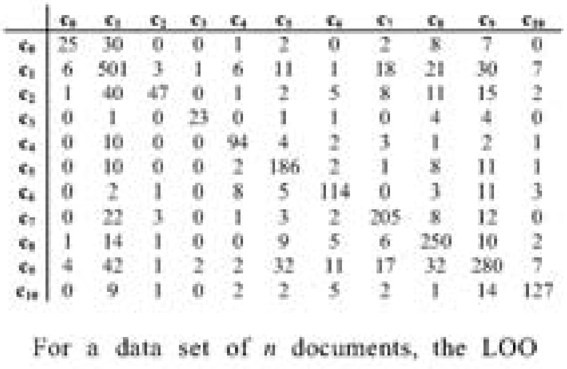


Fig. 4 Naive Bayes and MSVM classifiers train accuracy.

The experiments showed that the MSVM classifier outperforms the Naïve Bayes one as regards the train accuracy (Fig. 4). Chosen the MSVM classifier, its generalization capability has been tested using a *Leave One Out* (LOO) strategy.

Tab. 3 LOO results of the MSVM classifier.



	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
c_0	25	30	0	0	1	2	0	2	8	7	0
c_1	6	501	3	1	6	11	1	18	21	30	7
c_2	1	40	47	0	1	2	5	8	11	15	2
c_3	0	1	0	23	0	1	1	0	4	4	0
c_4	0	10	0	0	94	4	2	3	1	2	1
c_5	0	10	0	0	2	186	2	1	8	11	1
c_6	0	2	1	0	8	5	114	0	3	11	3
c_7	0	22	3	0	1	3	2	205	8	12	0
c_8	1	14	1	0	0	9	5	6	250	10	2
c_9	4	42	1	2	2	32	11	17	32	280	7
c_{10}	0	9	1	0	2	2	5	2	1	14	127

For a data set of *n* documents, the LOO

For a data set of *n* documents, the LOO strategy consists in *n* experiments where all the *n* examples of the data set are used to train the classifier, except a different examples at each run used to test. The *LOO accuracy* is the number of correct tests with respect of all the entire number of tests. The experiments produce a LOO accuracy for the MSVM classifier of 74.7%. Tab. 3 shows the details of the MSVM classifier predictions for individual classes: rows report true classes while columns report predicted ones, so that the entry of the element (c_i,c_j) represents the number of documents of class c_i classified in class c_j.

https://doi.org/10.23106/dcmi.952108070

8. Query construction using a controlled vocabulary

As discussed in Section 5 the second facility provided by the Portal architecture able to cope with semantic search requirements is represented by a guide able to help users in formulating a query effectively expressing the semantics of their information needs.

In the Portal architecture, users query a DC metadata index (Fig. 1).

This allows users to access the Portal index by two query modalities (2):

metadata-based document querying (MBDQ);

keyword-based document querying (KBDQ), combined with *category-based document querying* (CBDQ).

Case 1) Advanced search: in this case users submit a query filling in the fields related to DC metadata, particularly the selection of a value for *dc:subject* is mandatory. Terms inserted in each box are required to match terms contained within the associated DC metadata. The legal category selected in *dc:subject* makes the query more selective since the retrieval is focused on documents belonging to this legal category.

Case 2) Simple search: in this case users submit a query, filling an unqualified text box with keywords. Terms inserted in the unqualified text box are required to match terms within metadata or full text, if any. Moreover, to make the retrieval more focused the systems provides facilities to contextualise the query (CBDQ). Without query contextualisation, in fact, the retrieval can be less accurate: a document not containing query terms will not be retrieved, even if it is relevant to the user information needs.

Therefore, facilities on user side are desirable in order to expand a query and to match at its best the user information needs. In such a way a larger number of relevant documents can be retrieved, improving the recall of the retrieval. The query contextualisation according to a legal category can improve also the retrieval precision.

In the Portal architecture these facilities are provided at the user interface level using the DoGi legal specialized controlled vocabulary aiming at guiding users in effectively formulating a query.

Here below is how the user, in KBDQ modality, interacts with the DoGi vocabulary.

Terms composing the query are firstly matched against the legal controlled vocabulary, thus identifying the possible legal categories they are related to. Before the search task is executed, the user is requested to choose one legal category among those ones selected by the system, thus contextualising the query itself.

Then the search task is activated by matching terms against the content of the related DC metadata of documents.

Documents matching both terms and legal category are selected by the system.

In case no documents match the chosen terms, the system tries to expand the query pointing to narrower or broader terms, according to the DoGi controlled vocabulary. Such terms, combined with the legal category, are used to newly query the Portal index.

This procedure allows:

to retrieve relevant documents for the user information needs, even if they do not contain the terms chosen by the user, thus enhancing recall in retrieval;

to retrieve only documents containing query terms, but belonging to a specific legal category, thus enhancing the precision in retrieval.

As for the automatic classification procedure, a benchmark with predefined user queries is under development to provide a measure of the effectiveness of the retrieval, within the Portal architecture, in terms of precision and recall.

9. Conclusions

High quality retrieval of legal information is the main objective of a specialised gateway. In the legal literature Portal created by ITTIG an approach has been developed to achieve effective retrieval in terms of both precision and recall. These features are implemented for the retrieval of both data coming from structured repositories and of web documents. The designed project aims at providing a single point of access into disparate repositories where categories of law, as content of *dc:subject*, automatically generated for web resources, are the essential metadata to point to relevant legal literature documents improving precision in retrieval. Facilities in query formulation are given to the users through the exploitation of a legal controlled vocabulary, improving recall.

The Portal to legal literature has tried to face the limitations such as silence and irrelevance, by adopting a strategy based on the peculiarity of legal terminology and on semantic intellectual tools reflecting the richness of legal terminology and the structure of law.

References

1. E. Francesconi and G. Peruginelli. Integration between structured repositories and web documents. In *Proc. of the DC Conference* 2003, pp. 99-107
2. E. Francesconi and G. Peruginelli. Opening the

https://doi.org/10.23106/dcmi.952108070

legal literature Portal to multilingual access. In *Proc. of the DC Conference* 2004, pp. 100-107.

3. B. Dervin. From the mind's high of the user: the sense-making qualitative – quantitative methodology in J. D. Glazier and R.R. Powell (eds.). Qualitative research in information management, Englewood, CO: Libraries Unlimited, pp. 61-84, 1992.
4. T. Jacobson and D. Fusani. Computer, System, and Subject Knowledge in novice searching of a full-text, multifiles database. *Library & Information Science Research,* 1992, vo1.4 , n. 1, pp. 97-106.
5. N. Belkin, R. Oddy and H. Brooks. ASK for information retrieval: part I. Background and theory. *Journal of documentation*, 38 (2), pp. 61-71.
6. F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (1), 1–47, 2002.
7. C. Apté, F.J. Damerau, S. W.: 1994, Automated learning of decision rules for text categorization, ACM Transactions on Information Systems 12(3), 233–251.
8. S.T. Dumais, H. Chen, Hierarchical classification of Web content, in Proc. of ACM International Conference on Research and Development in Information Retrieval, pp. 256-263, 2000
9. Y. Yang and J.O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, in *Proc. of the Fourteenth International Conference on Machine Learning*, pp. 412-420, 1997.
10. J.R. Quinlan, Inductive Learning of Decision Trees, in Machine Learning 1, pp.81-106, 1986.
11. C. Bishop: Neural networks for pattern recognition, Oxford University Press, 1995.
12. F.V. Jensen, Introduction to Bayesian Networks, Springer-Verlag, 1996
13. C. Cortes, V. Vapnik, Support Vector Networks, Machine Learning 20, 1–25, 1995.
14. A. Passerini, Kernel Methods, Multiclass Classification and Applications to Computational Molecular Biology, Ph.d thesis, Università di Firenze, Italy, 2004.
15. V.N. Vapnik, Statistical Learning Theory, Wiley, 1998.
16. T. Joachims, A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, *in Proc. of the Fourteenth International Conference on Machine Learning*, pp. 143-151, 1997.
17. B. Schölkopf and A.J. Smola, Learning with Kernels, The MIT Press, 2002.
18. J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004
19. C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, Kluwer Academic Publishers, vol. 2, pp. 1-43, 1998
20. C. W. Hsu and C.-J. Lin, A comparison of methods for multi-class support vector machines}, *Trans. on Neural Networks*, 2 (13), pp. 415-425, 2002.
21. K. Crammer and Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines, Journal on Machine Learning Research, MIT Press, vol.2, pp.265-292, 2002.

https://doi.org/10.23106/dcmi.952108070