Identification of Web Information using Concept Hierarchies and On-line Updates of Concept Importance

Zhan Li[1], Ronald R. Yager[2] and Marek Reformat[1]

[1]thinkS[2]: thinking software and system laboratory
Electrical and Computer Engineering
University of Alberta, Edmonton, Canada, T6G 2V4

[2]Machine Intelligence Institute
Iona College, New Rochelle, NY 10801

Abstract—*The Internet users have to perform a lot of search to find web pages with relevant information. The paper proposes an approach for utilization of a hierarchy of concepts to perform identification of web pages in the environment of the Semantic Web. A user provides a hierarchy of concepts that can only partially "cover" their domain of interest. Ontologies related to that domain are used to instantiate the hierarchy with concrete information, as well as to enhance it with new concepts initially unknown to the user. A web page is checked against concepts from the hierarchy and activation levels of those concepts are aggregated using Ordered Weighted Averaging (OWA) operators that are part of the hierarchy of concepts. Mechanisms of aggregations embedded in OWA operators are determined using linguistic quantifiers and importance of concepts. The importance of concepts is constantly changing, and in order to automatically assign importance values and "keep track" of changes a new algorithm for Adaptive Assigning of Term Importance (AATI) is used.*

Keywords—*hierarchy of concepts, ontology, ordered weighted aggregation, text identification*

1 Introduction

The enormous growth of the WWW (World Wide Web) creates a serious challenge in discovering relevant information among all web documents. Those documents vary in quality of information they contain, as well as in their timeliness. It is not easy to find the most relevant documents when the repository is constantly changing.

In general, there are two approaches used for finding relevant web pages, i.e., "Information retrieval" (IR) [4] and "Text Categorization" (TC) [8]. IR retrieves documents based on a query using a "keyword-matching" mechanism without considering different meanings of keywords under different contexts. In TC, there is no query, and the knowledge about a given category is embedded in models developed using machine learning (ML) techniques based on documents representing the category.

In real world scenarios, people relay on Search Engines (SEs) to retrieve required information. SEs accept users' queries and return sets of related web documents applying one of two approaches: building a full-text index-based search (IR-like approach), like Google and Altavista; or constructing taxonomies with web documents populating categories (TC-like approach), such as Yahoo.

People identify documents as belonging to a specific category based on a set of keywords. A person looks at a document and "searches" for keywords and associated with them concepts related to a given category. Moreover, for a person all keywords and concepts are interconnected and constitute a network. Activation of one keyword or concept initiates activation of related concepts. Relations between them can be of different nature: some can "point" to more *specific* concepts, some can relate to concepts that are *parts of* the original concept, and some can relate to concepts that *contain* the original one. More keywords and associated concepts are found in the document more support is collected towards assigning this document to a category described by the keywords.

An introduction of the Semantic Web [1] has led to the increased popularity of ontology as a means of representing knowledge in a way that a human being as well as a machine can understand. An ontology is a formal, explicit specification of a shared conceptualization [2]. It is a set of well-defined classes that describe data models in a specific domain. Together with their individuals (instances of classes), ontologies work as knowledge characters to express individual facts [7]. This new representation of knowledge introduced to the web environment brings new possibilities of utilization of information [5,6].

A different format of knowledge representation has been proposed in [10]. The proposed format, called Hierarchy of Concepts (*HOFC*), represents concepts with atomic attributes or other concepts. As the result a graph-like structure is established where each vertex is a concept, and terminal vertices are attributes. The edges of the hierarchy represent relationships that define concepts with a set of other concepts and/or attributes. An *HOFC* can be used for representing any human-like structure of concepts, for example queries [10].

ISBN: 978-989-95079-6-8

The paper introduces a new idea for constructing a category identifier using an *HOFC* combined with ontology and an algorithm for automatic updating of concept importance. In a nutshell, a user provides a simple hierarchy describing their category of interest. This hierarchy contains concepts that are the most obvious for the user. The concepts are linked together via different types of relationships. The concepts that constitute the hierarchy are generic – they are just definitions of concepts. The specific pieces of information are extracted form ontologies and "attached" to the equivalent concepts of the hierarchy as attributes – in other words ontologies are used to populate the hierarchy. Relations existing between ontology classes are used to find new concepts that are added to the hierarchy as well. This provides a mechanism for dynamic enhancement of hierarchies.

The dynamic character of the proposed approach is evident in a process of determining importance of *HOFC* concepts. Concept importance changes accordingly to changes in the contents of web pages – newly posted documents and texts affect importance of keywords in their ability to "identify" a given category. In order to "follow" these changes we equipped our *HOFC* with *AATI* [3]. The *AATI* is an un-supervised algorithm with ability to determine and update concept importance values automatically.

2 Background

2.1 Hierarchy of Concepts

In a nutshell, an *HOFC* uses a notion of representing concepts with atomic attributes or other concepts. Edges of *HOFC* are of significant importance to the whole idea of *HOFC*. If we assume that a concept C_1 is defined, described by or related to other two concepts C_2 and C_3, then the hierarchy will have two edges connecting C_2 with C_1, and C_3 with C_1. The concept C_1 is called a higher-level concept, and C_2 and C_3 are lower-level concepts. This also means that activation of concepts C_2 and C_3 leads to activation of C_1.

*HOFC*s introduce a very important feature – the process of activation of higher-level concepts by lower-level concepts is fully controlled by a user. There are two controlling components: importance vector $\boldsymbol{M}$ and linguistic quantifier Q. The vector $\boldsymbol{M}$ is an indicator of significance of each lower-level concept in defining a higher-level concept. In other words, $\boldsymbol{M}$ determines a weight of each of the participating (lower-level) concepts in a process of identifying an activation level of higher-level concept. The linguistic quantifier Q guides the aggregation of lower-level concept activations levels. We can have such quantifiers as *some*, *most*, *at least half*, *about 1/3*, or *for all*. Both $\boldsymbol{M}$ and Q determine how activation levels of lower-level concepts should be combined using the Ordered Weighted Averaging (*OWA*) operator [9].

In a formal representation, the *OWA* operator, defined on the unit interval I and having dimension n, is a mapping F_w: $I^n \rightarrow I$ such that:

$$F_w(a_1,\ldots,a_n) = \sum_{j=1}^{n} w_j * b_j \qquad (1)$$

where b_j is the j^{th} largest of all arguments $a_1, a_2,\ldots, a_n$, and w_j is a weight such that w_j is in $[0, 1]$ and the sum of w_j is equal 1. If $\boldsymbol{W}$ is an n-dimensional vector whose j^{th} component is w_j and $\boldsymbol{B}$ is an n-dimensional vector whose j^{th} component is b_j then $F_w(a_1, a_2,\ldots, a_n) = \boldsymbol{W}^T\boldsymbol{B}$. In this formulation $\boldsymbol{W}$ is referred to as the *OWA* weighting vector and $\boldsymbol{B}$ is called the ordered argument vector.

2.2 Ontology

The most important aspect of the ontology used for Semantic Web applications is related to identifying two ontology layers: the ontology definition layer, and the ontology instance layer[1].

The *ontology definition layer* represents a framework used for establishing a structure of ontology and for defining classes (concepts)[2] existing in a given domain. A structure of ontology is built based on a relation *is-a* between classes. This relation represents a *subClassOf* connection between a superclass and a subclass.

The ontology definition contains descriptions of classes. Classes are defined using two types of the properties:

- *datatype property* – this type of property focuses on describing features of a class; this property can be expressed as values of such data types as boolean, float, integer, string, and many more (for example, byte, date, decimal, time);
- *object property* – this property defines other than *is-a* relationships among classes (nodes); these relationships follow the notion of Resource Description Framework (RDF) [11] that is based on a triple *subject-predicate-object*, where: *subject* identifies what object the triple is describing; *predicate* (property) defines the piece of data in the object a value is given to; and *object* is the actual value of the property; for example, the triple "John likes books" has "John" as subject, "likes" as predicate and "books" as object.

Once the ontology definition is constructed, it can be instantiated. The properties of classes are filled out with real data – values are assigned to datatype properties, and links to instances of other classes are assigned to object properties.

2.3 AATI

Knowledge about a specific category is understood as the importance of keywords describing a category. The keywords and their importance are used to identify documents that belong to the category. The primary characteristic of a new algorithm for *Adaptive Assigning of Terms Importance (AATI)* is that it updates weights of keywords – called here terms – at the same time it categorizes web documents. It means that a training phase is not required [3]. The proposed algorithm is able to handle changes of terms and their importance. Knowledge collected by *AATI* is constantly updated, and there is no need to restart the algorithm when terms and their importance change.

AATI works on two sets of terms. One set contains terms that represent a given category, which is called *target* category,

[1] According to the terminology adopted by the Semantic Web community, the term "instance" has been replaced by the term "individual". For the purpose of clarity, the term instance – similar to the term *instantiated* – is used throughout the paper

[2] The term class is used to distinguish ontology concepts from *HOFC* concepts.

ISBN: 978-989-95079-6-8

while the other set contains terms that do not belong to a *target* category. *AATI* updates importance of the terms via changing weights associated with the terms called Term Weights (*TW*). The importance of terms reflects their discriminatory power in identifying a *target* category. Each web document is evaluated by summing *TW* of all terms that the page contains. The sum is called Page Value (*PV*). Please note that *TW* and *PV* are recursively defined. That is, *TW* and *PV* depend on each other.
Let a set $\boldsymbol{P} = \{p_1,p_2,p_3,...,p_m\}$ denotes *PV* of web documents, where each element p_i denotes *PV* of the *i*th document. Let a set $\boldsymbol{T} = \{t_1,t_2,t_3,...,t_n\}$ denotes terms in web documents, where each element t_j represents *TW* of the term *j*. In *AATI*, *TW* defines *PV*:

$$p_i = \sum_{j=0}^{n} \frac{f_{ij}}{N_i} t_j \tag{2}$$

where N_i is the number of words on a page *i*; f_{ij} is the number of occurrences of term *j* on the page *i*. If a term *j* does not occur on page *i*, f_{ij} is zero.
At the same time, *PV* defines *TW*:

$$t_j = \sum_{i=0}^{m} \frac{f_{ij}}{N'_j} p_i \tag{3}$$

where N'_j is the number of occurrences of term *j* in the document set containing *m* documents. If a term does not occur in the page *i*, f_{ij} is zero.
Therefore in the language of linear algebra, the definition between *PV* and *TW* can be written as

$$\begin{matrix} P = A \cdot T \\ T = B \cdot P \end{matrix} \Rightarrow T = (B \cdot A) \cdot T \Rightarrow T = C \cdot T \tag{4}$$

Thus, $\boldsymbol{T}$ is the eigenvector of matrix $\boldsymbol{C}$ with the eigenvalue one. At the same time, because both the sum of each row in the matrix $\boldsymbol{A}$ and that of each row in the matrix $\boldsymbol{B}$ are one, the sum of each row in the matrix $\boldsymbol{C}$ is also one. So, the eigenvector *T* as the solution of (4) exists and is unique.
Therefore, we define the algorithm *AATI* to evaluate *TW* :

- ***Step1***: build a list of terms for a *target* category, and assign zero as *TW* for all terms;
- ***Step2***: take a new web document; if there is no more web documents, ***STOP***;
- ***Step3***: parse the document based on the list of terms;
- ***Step4***: for each term occurring in the document do one of the following:
  (a) If the term value is not zero (it means the term has already appeared in documents), take this value as *TW*;
  (b) If the term value is zero (the term has not been found in any documents), randomly generate a number between 0 and 1, and assume it as *TW* for this term;
- ***Step5***: Calculate the *PV* for this web document by Equation (2)
- ***Step6***: Update the *TW* of those terms found in the web document (3);
- ***Step7***: Normalize TW across all terms;
- ***Step8***: Go back to Step2.

3 HOFC as Keyword Structure for Text Identification

3.1 Description of Domain of Interest

Text categorization uses keywords as representation of a category, and a text identification process checks if a given text contains those keywords. If all or fraction of keywords (depending of the applied technique) are found in a given document then the document is considered as belonging to a category described by the keywords. Usually, all keywords are not equally important and they are associated with weights identifying their importance. The keywords are represented as a flat structure, and there is no indication that they are "related" to each other, and that some of them depend on others.
A human, as it has been already indicated, looks at a document and uses not only keywords but also related keywords to identify a category the document belongs to. The presence of one keyword in the document activates other connected keywords. More keywords are found in the document more support is collected towards assigning this document to a category described by those keywords.
This human-like approach to use a network of keywords describing a specific category can be implemented with an *HOFC* (Section 2.1). An *HOFC* defines a concept with a set of other concepts or attributes[3]. The relations between a defined (higher-level) concept and defining (lower-level) concepts and attributes represent different types of associations that are meaningful for a user. Therefore, a structure of *HOFC* can be treated as a user-defined description of a given category.
This analogy goes even further when we take a closer look at the process of activation of *HOFC* concepts. In a nutshell, an activation level of higher-level concepts depends on activation levels of lower-level concepts. In other words, activation of a higher-level concept emerges via activation of more specific lower-level concepts and attributes. A level of activation of higher-level concepts is determined via an aggregation operator *OWA* defined with a linguistic quantifier (Q) and an importance vector ($\boldsymbol{M}$).

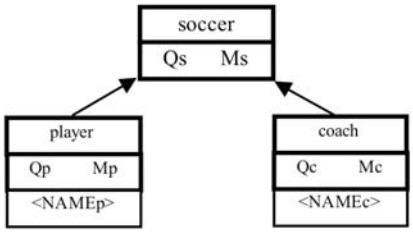

Figure 1: A simple structure of *HOFC* representing the category *soccer*

Fig. 1 shows a simple *HOFC* representing a *soccer* concept defined by two lower-level concepts: *player*, and *coach*. Each

[3] To ensure clarity it should be noted that among all phrases used in the paper the following relations hold: concepts are equivalent to classes; keywords, attributes and terms are equivalent to each other.

ISBN: 978-989-95079-6-8

of them has a single attribute: <NAMEp> and <NAMEc>, respectively. Concepts *soccer*, *player* and *coach* have their own linguistic quantifiers: *Qs*, *Qp*, and *Qc*, and importance vectors ***Ms***, ***Mp***, and ***Mc***. The vector ***Ms*** is of dimensionality 2, dimensionalities of ***Mp*** and ***Mc*** depend on a number of instances of each concept. Instances, that represent concrete pieces of information, are values of attributes <NAMEp> and <NAMEc>.

3.2 HOFC and Ontology

The application of *HOFC* to represent a user's view of the domain of interests brings a number of benefits. The *HOFC* allows us to incorporate different types of relations between concepts and attributes, and build quite complicated and well-interconnected nets of keywords/concepts describing the domain of interest.

However, users who provide *HOFC*s do not have to be experts in the domain. They can provide a very simple hierarchy – called a basic *HOFC* – that contains just a few concepts. This basic *HOFC* is treated as a starting point for building a more comprehensive *HOFC* better suited for identifying a category a given web page belongs to. The process of enriching the user-provided hierarchy is performed with ontology.

An ontology as proposed by the Semantic Web community (Section 2.2) represents a very attractive and powerful approach for knowledge representation. The richness of information that can be expressed with an ontology means that an ontology can be used to provide concrete and additional information for *HOFC*s. In the first case, we talk about instantiation of *HOFC*, and about enrichment of a hierarchy in the second case.

The process of instantiation can be explained with the help of a basic *HOFC*, Fig. 1. This hierarchy is "instantiated" with specific information obtained from the *soccer* category. The ontology is queried about *player*. The result – three instances of the class *player* – is presented in Fig. 2. The process of assigning values to *Q* and ***M*** is described in Section 3.3 and 4.

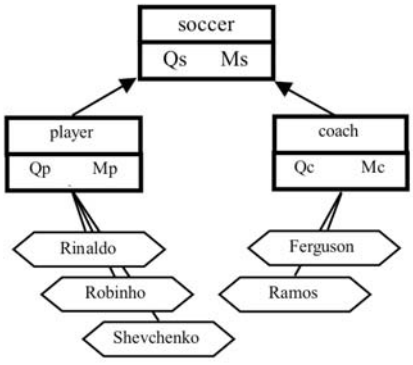

Figure 2: The HOFC of the category *soccer* with instances

An ontology can also be treated as a repository of new concepts that can be added to the concepts already existing in a *HOFC*. This enhancement can be twofold: in the form of addition of lower-level concepts, and in the form of providing concepts related in a "non-trivial" way to concepts already held in the hierarchy.

The concepts provided by a user are used to find corresponding classes in an ontology. The sub-classes of these classes are taken from the ontology and attached to the concepts of *HOFC* as new lower-level concepts. For an *HOFC* presented in Fig. 1, it would mean lower-level concepts: *forward*, *defender*, and *midfielder* are attached to the concept *player*.

Classes defined in an ontology possess object properties (Section 2.2). Object properties are definitions of "non-trivial" relations that exist between pairs of ontology classes. They are used to bring new concepts to an *HOFC*.

For example, the ontology class *player* has an object property *playsFor* that "binds" any instance of *player* with an instance of the ontology class *team*. If we assume that there is an instance of the class *player* called *Rinaldo*. *Rinaldo* has an object property *playsFor* that binds *Rinaldo* with an instance of the class *team – Manchester_United*, Fig. 3.

3.3 Text Identification process

A web page identification process corresponds to an "activation" of the top concept of an *HOFC*. The activation level of the top concept depends on activation levels of lower-level concepts and attributes. The activation of *HOFC* is determined in a bottom-up manner. Attributes at the bottom of *HOFC* are "matched" against words that appear in a web page. Presence of those words means activation of attributes, and activation of attributes means activation of upper-level concepts. This approach requires aggregation of activation levels of all attributes/concepts that define (are attached to) a concept. The aggregation process is performed using *OWA* (Section 2.1). The weights of *OWA* are determined based on *Q* and ***M***.

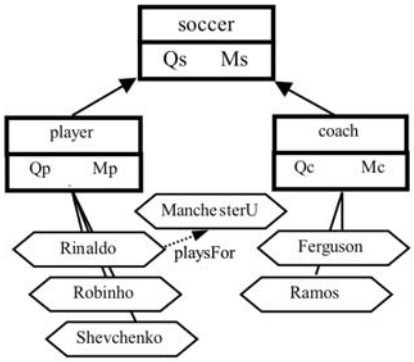

Figure 3: The *HOFC* of the category *soccer* with individuals and "enhancement"

Let $m_i \in [0,1]$ be a value associated with an attribute or concept a_i indicating its importance. If n concepts define a single higher-level concept then ***M*** of this concept is a n-dimensional importance vector $[m_1, m_2, \ldots, m_n]$, and the weighing vector ***W*** has to be calculated based upon both *Q* and ***M***. The first step is to determine the ordered argument vector ***B***, such that b_j is the j^{th} largest of all arguments $a_1, a_2, \ldots, a_n$. The arguments are numbers of occurrences of attributes in a

ISBN: 978-989-95079-6-8

text. Furthermore, we assume μ_j to denote the importance value associated with the attribute or concept that has the j^{th} largest value. Thus if a_5 is the largest value, then $b_1=a_5$ and $\mu_1=m_5$. The next step is to calculate the *OWA* weighing vector $\boldsymbol{W}$ using:

$$w_j = Q(\frac{S_j}{T}) - Q(\frac{S_{j-1}}{T})$$

where

$$S_j = \sum_{k=1}^{j} \mu_k \quad \text{and} \quad T = S_n = \sum_{k=1}^{n} \mu_k \cdot$$

So, S_j is the sum of the importances of the j^{th} most satisfied attributes/concpets, and T is the sum of all importances.
The linguistic quantifier Q is provided by a user at the time they provides their basic *HOFC*. During processes of enhancement, the values of Q for lower-level attributes and concepts are "inherited" from the upper-level concepts. In the case this is not possible a default quantifier *OR* is used.
The values of $\boldsymbol{M}$ representing importance of attributes and concepts of *HOFC* need to reflect a true significance of keywords in their capability to identify a category. The *AATI* algorithm capable of finding those values is used.

4 Importance of HofC Concepts

4.1 HOFC and AATI

Importance levels associated with each concept, keyword/attribute are considered as the knowledge about a domain of interest. Usage of the most accurate and relevant values is important for a good performance of the proposed *HOFC*-based approach. Therefore, the features of *AATI* (Section 2.3) make it a very good candidate for a process of updating and "tracking" changes in importance values of keywords. For the *AATI* algorithm, the importance values $\boldsymbol{M}$ are represented by *TW*.
The *AATI* algorithm modifies *TW* at the same time when it evaluates *PV* for each web document. This process is "controlled" by upcoming pages. If the content of stream of web documents changes, the *TW*s change too.
All concepts/keywords of *HOFC* are input to *AATI*. According to the algorithm their importance is randomly initiated but when time progresses and web pages are being processed – the values of *TW* are being updated and reflect strength of a connection of a term with a domain – strength of a term to "identify" the domain. These *TW* values are fed back to *HOFC* as $\boldsymbol{M}$ values.

4.2 Continuous Changes in Concept Importance

The ability of *AATI* to "follow" the web trends in popularity of a given keyword (popularity associated with a frequency of occurrence) can be illustrated with two examples. The first example represents a change in importance values when a source of web pages changes, while he second focuses on changes in the contents of web page over time.
Fig. 4 and 5 present examples of the adaptability of the *AATI* algorithm. Fig. 4 shows *TW* values of the term "Manchester United" (soccer ontology) as a function of a number of "processed" web pages. The first 1000 pages are from the BBC website (http://www.bbc.co.uk). The next 600 pages are from the CCN website (http://www.cnn.com). We see a decrease in the TW value (solid line after 1000 pages). If the next 600 were from the BBC website the values would follow the dashed line.

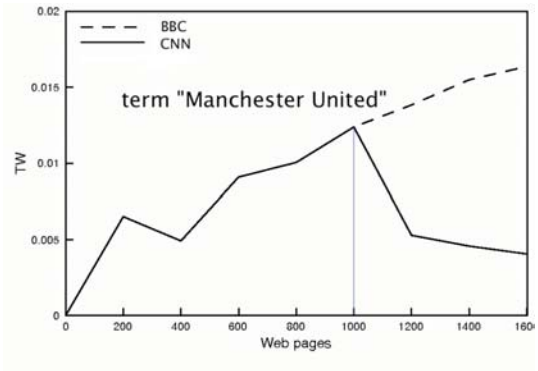

Figure 4: TW changes when the source of upcoming documents changes from BBC to CNN

The term "Manchester United" represents a name of a famous soccer club in England, and its *TW* goes up when the documents are downloaded form BBC. However, when the source of web pages changes and the upcoming documents are pages collected from the CNN site, and the term "Manchester United" is less popular. This decrease in popularity of the term "Manchester United" translates into decrease in its *TW*.
A different kind of change is seen in Fig. 5. The plots represent changes in importance levels when web pages collected at different moments in times are fed into the *AATI* algorithm. The first 2300 pages were collected from the BBC site in February 2007, while the last 2500 pages are from BBC site from January 2009. It can be seen that the importance levels for different terms behave differently: the importance of the player Ronaldo stays more or less at same level – indication that Ronaldo was and still is popular; the importance of the player Robinho increased and then slowly decreased, while the popularity of the player Shevchenko decreased.

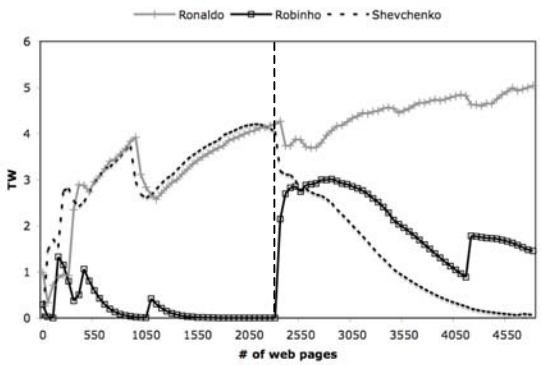

Figure 5: Changes of *TW* when the set of upcoming documents change from Feb07 to Jan09: for Ronaldo, Robinho, and Shevchenko

ISBN: 978-989-95079-6-8

5 Classification System

5.1 Architecture

An implementation of the proposed approach for web content identification requires integration of multiple components executing different tasks. A simplified architecture of a system performing the identification task is shown in Fig. 6. The system contains the following three parts: *HOFC* Preparation and Enhancement Unit, *AATI* Execution Unit, and Identification Unit. The *HOFC* Preperation and Enhancment Unit is responsible for setting up a basic *HOFC* representing user's query, adding new concepts from a domain ontology, and populating an enhanced *HOFC* with concrete pieces of information from the ontology. The *AATI* Unit is invoked on regular basis (based on time interval or a number of incoming web pages) to update values of importance of concepts. The Identification Unit checks the incoming web page against the concepts from the *HOFC*. This process starts with concepts located at the bottom of the *HOFC* – and it progresses towards the top. It is checked if the *HOFC* concepts are "activated". The activation level of concepts is the result of a number of their occurrences in a text, as well as Q and $\boldsymbol{M}$ (Section 3.3).
A prototype of the system has been developed in Java. The Protégé API is used to access ontology classes, their properties and instances. Moreover, API functions allow for creating a new enhanced hierarchy of concepts based on classes and individuals obtained from the domain ontology and the hierarchy of concepts given by a user.

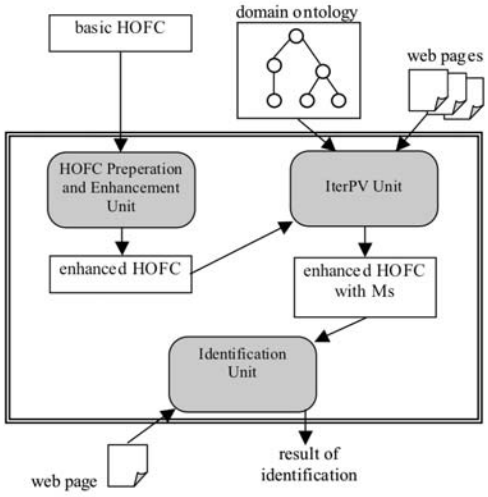

Figure 6: Text Identification System

5.2 Experimental Results

A simple example that illustrates application of the prototype system is presented. The *HOFC* used in the experiment is shown in Fig. 1. It is a very simple hierarchy representing the category *soccer*. A user is a novice to this category, and the only concepts that define soccer are *players* and *coach*. This *HOFC* is later instantiated and enhanced, Fig 3. The enhanced *HOFC* represents the following structure of concepts:

> The concept *soccer* is activated if there are "MOST" of *players* and *coaches*. The concept *players* is activated if there exist "OR" Ronaldo "OR" Robinho "OR" Shevchenko. The concept *coach* is activated when "OR" Ferguson "OR" Ramos are present.

This simple *HOFC* has been checked against a number of different web pages. In all cases the activation levels of the concept *soccer* reflected a human evaluation of pages. The activation levels have been calculated based on aggregations of activations of keywords – Ronaldo, Robinho, Shevchenko, Ferguson, Ramos – using simple linguistic quantifiers MOST and OR, and values of $\boldsymbol{Ms}$ provided by *AATI*.

6 Conclusions

The paper proposes a concept-based structure suitable for text identification. The method uses hierarchies of concepts as category identifiers. The hierarchies use *OWA* operators with linguistic quantifiers and concept importance to control conditions that have to be satisfied by lower-level concepts and attributes in order to activate a higher-level concept.
The concept importance levels are updated using the *AATI* algorithm. This algorithm is able to modify concept and attribute importance values in on-line fashion in an un-supervised manner.
A prototype of the system implementing the proposed approach is briefly described in the paper.

Reference

[1] T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web, *Scientific American*, 284: 34-43, May 2001.

[2] T.E. Gruber, A translation approach to portable ontology specifications, *Knowledge Acquisition*, 5: 199-220, 1993.

[3] Z. Li and M. Reformat, An Iterative Algorithm for Text Categorization, submitted to *Information Processing Letters*.

[4] C. Manning, P. Raghavan, and H. Schutze, *Introduction to information retrieval*, Cambridge University Press, New York, NY, 2008

[5] T. Martin, Searching and smushing on the semantic web-challenges for soft computing, in M. Nikravesh, B. Azvine, R. R. Yager, & L. A. Zadeh (Ed.), *Enhancing the Power of the Web*, Heidelberg: Springer, 167-188, 2004.

[6] E. Sanchez, and T. Yamanoi, Fuzzy Ontologies for the Semantic Web, in *Flexible Query Answering Systems*, 691-699, 2006.

[7] F. Scott, W.D. Lewis, and D.T. Langendoen, An ontology for Linguistic Annotation 1-2, in Proceedings of *14th Innovative Applications of AI Conference*, Edmonton, Canada, 11-19, 2002.

[8] F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Survey(CSUR)*, 34(1): 1-47, 2002.

[9] R.R. Yager, On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18: 183-190, 1988.

[10] R.R. Yager, A Hierarchical Document Retrieval Language, *Information Retrieval*, 3: 357-377, 2000.

[11] RDF: http://www.w3.org/RDF/

ISBN: 978-989-95079-6-8