Information Extraction, Search, Interaction and Collaboration on the Web in Mexico

J. Alfredo Sánchez[1], Edgar Chávez[2], Manuel Montes[3]
[1]*Universidad de las Américas Puebla,* [2]*Universidad Michoacana de San Nicolás de Hidalgo,*
[3]*Instituto Nacional de Astrofísica, Óptica y Electrónica*
j.alfredo.sanchez@gmail.com, elchavez@fismat.umich.mx, mmontesg@inaoep.mx

Abstract

Web research in Mexico has been addressing issues related mainly to search mechanisms, information extraction, and mediating user interaction and group collaboration. In this paper we provide an overview of representative projects in the area and present a sample of recent advances by research groups in Mexican institutions. These include initiatives aimed to exploring extraction techniques that regard the web as a corpus, indexing and categorizing multimedia web contents, and designing user interfaces for visualizing web-accessible collections as well as environments for synchronous and asynchronous web collaboration.

1. Introduction

A significant amount of research in computer science today has some sort of relationship with the web: the focus may be at its core, or the web may be used as the infrastructure for interactive or collaborative applications, or it may be regarded as a vast corpus upon which search and retrieval mechanisms are tested and evaluated. Computer science research in Mexico is not an exception. In this paper, we sample what has been explored and accomplished in recent years by various research groups in Mexico that rely on the web as a context for their work.

A look at previous LA-Web conferences can provide some insight into the emphases of the web research efforts that have been undertaken in Mexico, essentially in the categories mentioned above. In the area of web search and retrieval, Medina et al. [15] discussed initial results of a applying a document retrieval method to distributed collections available on the web via the OAI Protocol for Metadata Harvesting (OAI-PMH). Their method was based on the construction of ontologies using document clustering techniques. Further progress in this effort was reported the following year [14], when methods were proposed for ontology maintenance via collaborative rewriting and revision.

The role of context in improving the results of web search mechanisms is investigated by Ramírez and Brena [18]. They propose the notion of "semantic contexts" as clusters in keyword spaces and report experimental results that show how semantic contexts may contribute to refining web search processes. Silva and Favela [24] also investigate how the context of queries may impact the performance of web search engines. In particular, they describe and report experimental results of using a search tool that helps users find health information on the web that is relevant to their current conditions by using personal health records. Another effort that is aimed to better characterize web contents is discussed by García et al. [4]. Their goal is to categorize text documents by measuring the amount of information or entropy in the text. The entropy is computed as an empirical distribution of words in a given text.

Work has also been conducted which takes advantage of the web as an infrastructure for distributed computing. Thus, Téllez et al. [26] discuss an architecture based on web services and a system designed to facilitate transparent and translucent remote object access. Contreras and Hernández [2] apply the notion of shared ontologies to negotiation problems that arise in a web-based environment in which heterogeneous agents that rely on different languages. They describe a methodology for integrating and publishing descriptions of negotiation primitives in a shared ontology.

Finally, in the realm of interactive and collaborative applications, Sánchez et al. [19] describe an environment that supports personal and public annotations on web documents, as well as the generation of index cards intended to help organize

and expedite scholarly work on the web. Martínez-Ruiz et al. [13] discuss a method for designing web user interfaces by applying an iterative series of XSLT transformations that translate an abstract interface model to an interface that is coded for a specific platform. Given the widespread use of the web for supporting distance education and e-learning, the work by Vázquez and Ostróvskaya [27] may be of relevance for new projects in the field. They analyze open standards that are commonly used when constructing learning objects, and show how the results of that analysis are applied in making appropriate choices in the context of a specific project for developing voice-enhanced learning objects available on the web.

Naturally, web research has continued in various institutions throughout Mexico. In what follows, we present with more detail a sample of recent advances in three major categories: information extraction, indexing and searching, and user interfaces and collaboration environments.

2. Extracting information from the web

Nowadays, one of the most popular research topics in computational linguistics and automatic text processing is the extraction of information from the web or, in other words, the usage of the *web as corpus* [12]. In particular, in Mexico there have been some interesting efforts related to the use of the web for the automatic construction of domain-specific ontologies [16], training sets for text classification tasks [6, 7], and language models for speech recognition [28]. The following sections give a brief overview of these works.

2.1 Using lexical patterns to extract hyponyms

Linguistic resources such as ontologies have a broad range of applications in computational linguistics and automatic text processing. They provide significant knowledge about languages but are expensive to build, maintain and extend. As a result, their usefulness is still restricted to certain domains or specific applications. In order to overcome this problem, recently many researchers have been working on semiautomatic methods for their construction. In particular, there is a special interest in the extraction of synonyms, antonyms and *hyponyms* from free text documents.

An example of these works is the one presented in [16], which describes a pattern-based method for the automatic acquisition of hyponyms (*is-a* relations) from the web. This method consists of the following three main modules:

Pattern discovery. It focuses on the discovery of lexical extraction patterns from the web, such as "*the HYPONYM is a HYPERONYM that.*" Its objective is to capture most writing conventions used to introduce a hyponym relation between two words.

Instance extraction. In this module, the discovered patterns are applied over a target document collection in order to locate text segments that presumably contain an instance of the hyponym relation. The result is a set of candidate hyponym-hypernym pairs such as *diamond-stone* and *BBVA-bank*. Nevertheless, it is also possible to discover incorrect instances such as *privatization–stone* (from the snippet "the privatization is one of the stones of market development of regional economy…").

Instance ranking. It evaluates the confidence of the extracted instances to belong to the hyponym relation. Its purpose is to rank the instances in such a way that those with higher probability of being correct are located at the very first positions. This evaluation is based on the idea that pertinent instances are extracted by different patterns, and that valuable patterns allow extracting several pertinent instances. In particular, it considers an iterative evaluation process where instances' confidence values are calculated based on patterns' confidence values, and vice versa.

An evaluation of this method was conducted using texts in Spanish. It considered a set of 25 seed instances corresponding to five different concepts. In the end, it enabled the construction of a corpus of 12,500 segments expressing the hyponym relation and extracted, with a precision higher than 75%, 851 candidate hyponym instances: 193 related to bank, 307 to disease, 9 to feline, 226 to profession, and 116 to stones.

2.2 Improving text categorization using the web as corpus

Text categorization, the assignment of free text documents to one or more predefined categories based on their content, has emerged as a very important component in many information management tasks. Its more successful approaches consider statistical and machine learning techniques. A major difficulty with these techniques is that they commonly require a great number of labeled examples (training instances) to construct an accurate classifier. Unfortunately, because a human expert must manually label these examples, the training sets are extremely small for many application domains. In order to overcome this

problem, recently many researchers have been working on *semi-supervised learning algorithms*.

In particular, [6] proposes a new method for semi-supervised text categorization. This method differs from previous approaches in that it does not require a predefined set of unlabeled examples; instead, it considers the automatic extraction of related untagged data from the web. This method consists of two main processes:

Corpora Acquisition. It focuses on the automatic extraction of unlabeled examples from the web. In order to do this, it first constructs a number of queries by combining the most significant words for each class, and then, using these queries, it looks at the web for some additional training examples related to the given classes.

Semi-supervised learning. Its purpose is to increase the classification accuracy by gradually augmenting the originally small training set with the examples downloaded from the web. In this case the selection of the examples is accomplished by an ensemble of SVM and Naïve Bayes classifiers.

Experimental results on a set of newspaper articles about natural disasters demonstrated the viability of the proposed method; using less than ten labeled examples per class it was possible to achieve a classification accuracy of 97.5%, outperforming baseline results by almost 30%. Furthermore, the method was also evaluated in a task of non-thematic text categorization [7], particularly, in a corpus of 353 poems written by five different authors. In this case, the preliminary results were very interesting, since they demonstrated that it is feasible to extract useful examples from the web for the task of authorship attribution. In fact, our intuition suggested the opposite: given that poems tend to use rare and improper word combinations, the web seemed not to be an adequate source of relevant information for this task.

2.3 Tuning task-specific language models through web data

The language model is an important component of any speech recognition system. Its purpose is to reduce the search space in order to accelerate and improve the recognition process. In particular, language models allow describing the use of words (mainly their common combinations) in a specific domain or task.

The generation of a language model requires the construction of a *large training corpus* that contains the greatest number of contexts for each word. The construction of this corpus is not a simple task since written texts do not represent adequately many phenomena of spontaneous speech. In order to alleviate this problem, [28] proposes the use of web documents as data source. This proposal was based on the fact that many people around the world contribute to create the web, and therefore, that most of its documents comprise informal contents and include many everyday as well as non-grammatical expressions used in spoken language.

In particular, the method presented in [28] faces the problem of enlarging a given small task-specific corpus (called reference corpus). It considers the following main steps:

Corpus acquisition. First, it employs the vocabulary from the reference corpus to gather from the web a large number of documents. Then, it selects from the downloaded documents the minimal blocks (word sequences) containing only words from the task vocabulary. The set of minimal block is defined as the new corpus.

Lexical analysis. It is clear that the terms and expressions used in real dialogs considerably differ from those occurring in texts. For instance, we can expect that the frequency of pronouns and verbs in the first and second person is not similar between a dialog among people and a written text. Therefore, the aim of this analysis is to find those words having very different frequencies in both corpora (i.e. between the new corpus and a reference corpus). The identified critical words can be over or under represented in the new corpus in relation to the reference one.

Corpus enrichment. The purpose of this process is to reduce the identified weaknesses of the new corpus. In order to do that, this process enlarges the new corpus by adding several copies of a selected set of phrases (containing the critical words) from the reference corpus.

The evaluation of the proposed method was carried out in the domain of kitchen design. The results obtained indicate, on the one hand, that the method could considerably increase the size of the training corpus (from 27,459 lexical forms to 27,224,579), and on the other hand, that the enrichment process allowed improving the quality of the original web-based corpus by more than 180 perplexity points.

3. Searching the web

Even assuming the huge amounts of data held by the web were properly stored and indexed, a challenge would remain to organize these repositories in a human-friendly fashion. Ranking (e.g. based on the PageRank algorithm) works well when searching with keywords, making it possible for the user to find

relevant web pages satisfactorily. Ranking algorithms are successful mainly because the link structure of the web is used as an external, static way to rank web pages containing the keywords being searched. Among the millions of pages containing the searched keywords, ranking algorithms provide a meaningful way to organize them. However, when browsing web pages that have not been previously cross-referenced or when looking for multimedia units such as songs, movies or photographs, the interaction and search models are completely different. In those cases, a (piece of the) object being searched is typically presented to the system, and a collection of documents or objects in decreasing order of similarity to the (piece of the) object being searched for is expected in return. This query-by-example kind of model can be stretched to encompass most pattern recognition tasks.

But the model is not as successful for multimedia web objects as it is for searching text on the web for two main reasons: firstly, there is not a clear understanding of how visual or auditory resemblance can be mapped to a mathematical formula or an algorithm, even though lots of attempts have been made to make progress in this respect. Secondly, even if we assume an opaque mapping that enables meaningful retrieval, the search will not scale well, since a nearly sequential scan of the object collection would need to be performed due to the so-called curse of dimensionality. It should be noted that text retrieval in this context is not very different from image, sound or video retrieval.

A group in Mexico has been making significant progress in defining meaningful similarity functions and generating scalable indexes for multimedia object retrieval which are applicable for efficiently searching multimedia web collections. With respect to similarity functions, promising results have been produced by using the entropy of documents and the entropy of audio or video signals. The technique at a glance consists of decomposing the multimedia object into overlapping parts, and then computing the amount of information contained in each part. The amount of information seems fundamental as a building block for mimicking human-perceived similarities.

Applications such as text categorization and multimedia identification will be of use for only a few thousands or millions of entries if a scalable index is not developed. Currently, the best indexing techniques for multimedia or general similarity functions are carried out by using metric indexes. The research group is heavily involved with the development of metric indexing techniques that are both scalable and flexible. The first Similarity Search and Applications Workshop was held in 2008 in Cancun, Mexico[1].

In what follows we provide further details of the similarity functions that have been developed based on the amount of information for two specific cases: text categorization and audio retrieval.

3.1 Text categorization

Two main approaches have been followed for classification for large collections of text documents: supervised and unsupervised. Text categorization is a supervised classification problem which consists of assigning a predefined category to text documents (e.g. politics, economics, sports, etc.). Document clustering consists of automatically segmenting a text collection to search for interesting associations; this is unsupervised classification. Both of them are classic problems in machine learning. A properly selected similarity measure between text documents can help improve these and other related techniques. A novel approach to define document similarity is to focus on the amount of information (the entropy measure) of the text. This measure is interesting because it does not depend on any external factors. Since the amount of information can be computed with local statistics alone, we only need the empirical distribution of words in the text.

In [4] a system is described that receives a manually segmented collection of documents in diverse categories. For each category a separate empirical distribution of words is computed, and these empirical distributions are used for categorization purposes. The computed entropy of a test document is maximal for the correct category. For example, the computed entropy of a *sports* document using the *politics* empirical word distribution is lower than the computed entropy for the same document with the *sports* empirical word distribution. The proposed text categorization approach is simple, easy to code and does not need any training time (aside from histogram computations). The classification time is linear on the size of the document and the number of document categories, which implies it is faster to compute. The proposed method has been shown experimentally to be equivalent to or better than state-of-the-art classifiers.

[1] A comprehensive guide to scalable metric indexing can be found at http://www.sisap.org.

3.2 Audio fingerprinting

Humans can identify, without any difficulty, two heavily degraded versions of the same audio sample. The audio signals of a lossy compressed audio file (e.g. MP3) and the original WAV file can be really different in the time domain, when we *see* the energy of the signal, and in the Fourier domain, when we *see* the frequency components. Lossy compression models the human hearing inability to perceive masked sounds (e.g. a very loud beat near a high pitched note). The ultimate example of this grouping is speech recognition. We are able to distinguish a word even when the signal-to-noise ratio is negative (when the noise is louder than the signal); for example in a party, a stadium or a night club. This behavior has puzzled the scientific community for several decades.

Ibarrola et al. [8, 9] produced a novel proposal to identify heavily degraded versions of audio signals. We can explain the technique through a realistic analogy. When listening to the radio the music we hear is demodulated in either amplitude or frequency (AM or FM, respectively). On the broadcasting end the music is convoluted in the time or frequency domains with a carrier signal. On the radio receiver end, the signal is demodulated to restore de original music. The innovative idea introduced by Ibarrola et al. [8, 9] is to demodulate the amount of information carried by the sound signal. The time-domain entropy signature (TES) is computed directly from the energy of the signal, whereas the multi-band spectral entropy signature (MBSES) is computed in the Fourier domain.

The technique excels in recognizing a song among thousands given a degraded excerpt of only five seconds. The technique also has been used for broadcast monitoring and cover song identification (when the same song is performed by different groups/artists).

4. Interacting and collaborating on the web

The vastness of the information space presented by the web also poses both opportunities and challenges for user interfaces. One of the research areas addressed by some of the projects discussed below is that of finding mechanisms that may lighten the cognitive overload and disorientation problems faced by users when exploring the web and seeking relationships among web documents. A complementary area takes advantage of the large number of users that access information and applications via the web, thus generating opportunities for synchronous and asynchronous collaborative work.

4.1 Visualizing web-accessible collections

Recent work in the area of user interfaces in Mexico has produced a novel approach for visualizing large collections that are distributed among multiple institutional web-accessible repositories. This approach has been termed *star-fish*, as it combines the method known as *starfields* with *fisheye* views.

Starfields map large information spaces to a two-dimensional grid, in which small dots stand for elements which exhibit attributes that are mapped to the horizontal and vertical axes [1]. Fisheye views rely on a distorted graph that places the main point of interest in an information space as the focus, which is magnified and shown in detail. Less relevant information elements appear slightly compressed further away from the focus, thus providing context for the viewer [3]. In Star-fish, starfields are enhanced with fisheye view functionality that allows users to keep track of the context of objects of their interest while visualizing very large collections of documents in distributed digital libraries. This is accomplished through successive refinements of starfield representations in which areas of interest are magnified whereas surrounding areas are minimized. The resulting interface also makes use of iconic representations in starfields for differing types of collections, as well as colored variants for representing the institutions that hold the collections. The approach has been implemented on top of the Open Network of Digital Libraries (ONeDL)[2], a federation of digital libraries enabled by the OAI-PMH protocol. Details of a prototypical implementation of Star-fish and results from initial user evaluation are reported in [23]. Work in progress in this area involves the visualization of semantic relationships among web documents.

Another effort, aimed to assist users in detecting and visualizing relationships among web documents, is the Center for Counter Plagiarism (CCP). This is a tool designed to allow users to provide documents for verifying possible plagiarism instances with respect to the web *docuverse*. Documents to be verified are divided into sentences and sent to several engines that retrieve similar matches. A user interface presents analyzed sentences graphically catalogued as *copied*, *similar* or *original,* using a color representation resembling that of a traffic light (red, yellow and green, respectively). Representations of increasing levels of detail are generated, thus allowing users to analyze documents from a high-level overview involving percentages, to a fine-grained detail of every

[2] http://www.onedl.org.mx

sentence in the document. This also allows the user to decide whether or not a more detailed investigation is required. For example, color dispersion in a document representation can be good indicator for detecting plagiarism: several red or yellow sentences in a row may correspond to a sizable chunk of text that has been copied from the same source. Details of the interface of CCP and experimental results have been reported in [17].

4.2 Fostering collaboration on the web

Advances related to collaboration on the web are roughly organized into asynchronous and synchronous collaboration projects.

4.2.1 Asynchronous collaboration

Tagging (or social bookmarking) clearly has become a ubiquitous mechanism that supports both personalized and collaborative organization of varied information spaces on the web. The success of social bookmarking has prompted a lot of work aimed to investigate, for example, how tagging occurs, how tag collections are structured, how close folksonomies are to formal classifications generated by specialists, or how tags can be used to enhance information retrieval.

Recent work conducted in Mexico has introduced the notion of induced tagging [22], a kind of social bookmarking with two key characteristics: (1) a well-defined group of participants are knowledgeable on the available resources and the background of the user community; and (2) tagging is required as part of that group's regular responsibilities as a reference team. The concept of induced tagging has been proposed to take advantage of the shift that is occurring in the role of information professionals, particularly reference librarians. In the process of helping users, staff and users often discover resources that might be useful for supporting current or future tasks. It should be possible for information experts to bookmark those resources and share their findings with their colleagues and the entire community they serve. Although all users are encouraged to tag, having a specialized group that does altruistic tagging continuously and applies tags consistently for extended time periods as part of their job, addresses concerns on the advantages of controlled vocabularies as well as incentive issues. Moreover, given the familiarity of the information experts with the needs of their user community, schemes can also be devised to facilitate the generation of personalized recommendations that are based on the tags assigned to relevant resources.

Induced tagging is being explored with the implementation and deployment of REC, an Ajax-based platform that provides a toolbar that can be added to a web browser so users may label resources in a minimally disruptive manner while they navigate around the web. Additionally, users may manage tags and request recommendations using the main REC interface, which is also web-based. Over the period of about six months, six information experts have been tagging resources using REC in addition to their regular duties, which include assisting users at the reference desk and via a virtual reference environment. Students also use REC as part of project assignments in a first-year college course.

Collections provided by different vendors typically are available through highly heterogeneous interfaces. When tags are assigned to those documents by the information experts (or other users), REC becomes a uniform interface that establishes tacit links that are used to discover resources comprised by diverse collections that otherwise would go unnoticed. Participating users have been able to fetch documents carefully selected by information experts, by just following links recommended by REC, and remaining essentially unaware of the specific collections that contain the documents. Tag clouds have resulted also in a discovery mechanism, as they invite users to explore relevant tags and their associated resources.

The combination of social networks and induced tagging can provide an even more powerful environment for exploiting information spaces. Not only are users familiar with various social networking systems, but they also are keen on extending their network of contacts, which could just as well include information experts that could help them accomplish academic endeavors without having to leave their social environment of choice. A version of REC was developed which can be added as an application of Facebook, one of the most popular social networking systems. Users of Facebook who install REC can obtain or give recommendations within the same familiar environment. Registration as a REC user is required to enable tagging only the first time a user tags resources. Both versions of REC share the same databases, so tags and recommendations managed by either of them are equivalent.

4.2.2 Synchronous collaboration

Making applications available via the web has promoted increased portability among hardware and software platforms, thus making it possible for users to interact with other local or remote users regardless of the diversity of their physical facilities, operating systems, or web browsers. Progress has been made in Mexico in building web-accessible applications that facilitate collaboration for users of large interactive displays, portable computers and mobile devices.

One recent development in this area is a web-based software environment named KBoard, which is aimed at supporting knowledge capture in the context of group activities mediated by interactive surfaces. KBoard implements the notion of workspaces that can be built, inter-linked and navigated as discussion progresses. Also, it provides alternative dynamic QWERTY and pie-menu soft keyboards that speed up text input, which is a key component of knowledge reification. The results of its preliminary evaluation indicate that KBoard has potential to facilitate knowledge capture, and provides a toolset to explore more fully the applications of large interactive surfaces. Details on KBoard have been reported in [20].

Another recent effort has produced an environment called STRATA, which combines the notion of telepointers with that of transparent annotation layers to foster effective collaboration. Telepointers are representations of users' cursors in a remote environment; they have been widely used in groupware systems [5], given their ability to convey information such as embodiment, gestures, and coordination among group members. By using telepointers in STRATA, both local and remote users are represented ostensibly and group awareness is achieved. STRATA introduces a novel technique for handling individual annotations based on transparent layers, which can be activated or deactivated to prevent display clutter and to support coordination and communication among group members. STRATA aims to support collaborative work in the context of multimedia collaboration rooms, in which heterogeneous platforms include large interactive displays, laptop or desktop computers and mobile devices. The main kinds of activities that are supported are those that involve discussion of materials presented by one of the group members who acts as facilitator or instructor. Details of the design of STRATA as well as results from initial use are reported in [21].

5. Perspectives

The web provides an appealing context for research in computer science and information technologies. In Mexico, there is an active community conducting web-related research at practically all levels of abstraction: from infrastructure and representational issues, to middleware and communications software, to applications and collaboration environments.

Work is expected to continue in all fronts and the web research community will continue to grow. The developments described in this paper are active and open issues are being explored by ongoing projects. Thus, for example, future work in information extraction includes probing the use of lexical patterns for extracting homonymy and synonymy relations, as well as applying web-based text-categorization methods to tasks such as word sense disambiguation and named entity recognition.

In the area of web indexing and searching, work will focus on efficient unsupervised classification of text, considering that the word distribution must be recomputed for each possible grouping. Another interesting problem being explored is genre and mood classification using fingerprinting techniques to convert a signal processing problem into a text problem. Audio signals are converted to text by using entropy-based audio fingerprint techniques as described previously [10, 11]. Ongoing work is extending, with excellent results, entropy-like techniques to the photography and video domains. Also, an open source suite for multimedia indexing and searching, referred to as the NATIX project, is being developed. NATIX will provide services both to the scientific community and the general public. For the general public, the project will provide free Robust Digital Object Identifier (R-DOI), so users will be able to compute the MBSES of an excerpt of a song (or a picture or a video), then send it to a web service, and the system will return a unique identifier that may be used to index the digital object metadata in a separate application[3].

Finally, in the area of web-mediated interaction and collaboration, experimentation and usability studies of visualization interfaces and collaboration environments have led to new designs and new prototypical implementations. Since many of the software prototypes are freely available on the

[3] NATIX services and related source code will be available at http://www.natix.org.

web[4], feedback from an extended community is also expected to inform the research being conducted. The convergence of user interfaces and information retrieval techniques is also generating relevant web research. For example, novel approaches to visual information retrieval (VIR) are being explored by a group in Mexico. These approaches cover a wide range of methods and techniques for content-based analysis of multimedia information. Some are based on extracting wavelet coefficients and computing similarity between images. A particular direction of this work is based on the combination of color-shape analysis of objects in images with their indexing based on textual descriptions. The principal goal of this technique is applying Two Segments Turning Function (2STF) and the "Star Field" (SF) approach for efficient processing and matching of invariant to spatial variations [25].

The web has an enormous potential to improve the accessibility of vast information repositories, and also to increase the presence of communities and societies in a global context through the production of original contents that promote and preserve our countries' cultures and values. It is a grand challenge for computer science researchers in Mexico and Latin America to use the web to empower our region in the emerging knowledge society.

References

[1] Ahlberg, C., Shneiderman, B. Visual Information Seeking: Tight coupling of dynamic query filters with starfield displays, *ACM CHI '94 Conference Proc.* (Boston, MA, April 24-28, 1994), 313-317.

[2] Contreras, M. C. and Hernández, J. C. Ontology Solution for Communicating Heterogeneous Negotiation Agents in a Web-based Environment. In *Proc. of the Fourth Latin American Web Congress*. LA-WEB. 2006. IEEE Computer Society, Washington, DC, 59-66.

[3] Furnas, G. W. Generalized fisheye views. In *Proc. of CHI '86*, 16-23. ACM Press (1986).

[4] García, M., Hidalgo, H., and Chávez, E.. Contextual Entropy and Text Categorization. In *Proc. of the Fourth Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC, (2006), pp. 147-153.

[5] Gutwin, C., & Penner, R.: Improving interpretation of remote gestures with telepointer traces. In *Proc. of the 2002 ACM conference on Computer Supported Cooperative Work*, 49-57 (2002).

[6] Guzmán-Cabrera, R., Montes-y-Gómez, M., Rosso, P., Villaseñor-Pineda, L. Improving Text Classification using Web Corpora. *5th Atlantic Web Intelligence Conference*, AWIC 2007. Advances in Soft Computing, Num. 43, Springer (2007).

[7] Guzmán-Cabrera, R., Montes-y-Gómez, R., Rosso, P., Villaseñor-Pineda, L. A Web-based Self-training Approach for Authorship Attribution. *6th International Conference on Natural Language Processing*, GoTAL 2008 (*Forthcoming*).

[8] Ibarrola, A. C., Chávez, E. A Robust Entropy-Based Audio-Fingerprint International Conference on Multimedia and Expo (ICME 2006), Toronto, Canada.

[9] Ibarrola, A. C., Chávez, E. On Musical Performances Identification, Entropy and String Matching. *Mexican International Conference on Artificial Intelligence* MICAI (2006) Tlaxcala, México.

[10] Ibarrola, C. A., Chávez, E. Robust Audio-Fingerprinting with Spectral Entropy Signatures, (2008). *Submitted*.

[11] Ibarrola, C. A., Chávez, E. Matching Musical Performances Using String Processing Techniques With Stable Features, (2008). *Submitted*.

[12] Kilgarriff, A. and Grefenstette, G. Web as Corpus. *Computational Linguistics*, Vol. 29. No. 3 (2003).

[13] Martínez-Ruiz, F. J., Arteaga, J. M., Vanderdonckt, J., González-Calleros, J. M., and Mendoza, R. A first draft of a Model-driven Method for Designing Graphical User Interfaces of Rich Internet Applications. In *Proc. of the Fourth Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC (2006), 32-38.

[14] Medina, M. A., Chávez-Aragon, A., and Chávez, R. O. Construction, Implementation and Maintenance of Ontologies of Records. In *Proc. of the Fourth Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC (2006) 67-73.

[15] Medina, M. A., Sánchez, J. A., Ostróvskaya, Y., and Brisaboa, N. R. OntoSIR: An OAI Service for Multi-Collection Document Retrieval Based on Ontologies of Metadata Records. In *Proc. Third Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC (2005), 111-114.

[16] Ortega-Mendoza, R., Villaseñor-Pineda, L. and Montes-y-Gómez, M. Using Lexical Patterns to Extract Hyponyms from the Web. *Mexican International Conference on Artificial Intelligence* MICAI 2007. Aguascalientes, Mexico, November. Lecture Notes in Artificial Intelligence 4827, Springer, (2007).

[4] Star-fish, CCP, REC, KBoard and Strata are available for use or download from the Laboratory of Interactive and Collaborative Technologies (http://ict.udlap.mx).

[17] Paredes, R. G., Sanchez, J. A., and Razo, A. Drawing the Line between Fair Use and Plagiarism for Digital Documents. In *Proc. of the Eighth Mexican international Conference on Current Trends in Computer Science*. ENC. IEEE Computer Society, Washington, DC (2007), pp. 113-122.

[18] Ramírez, E. H. and Brena, R. F. Semantic Contexts in the Internet. In *Proc. of the Fourth Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC (2006), pp. 74-81.

[19] Sánchez, J. A., Flores, L. A., Kirschning, I., and Ostróvskaya, Y. Supporting Web-Based Scholarship Through Index Cards and Annotations. In *Proc. of the Third Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC (2005), 54-59.

[20] Sánchez, J. A., Garnica, M., Valdiviezo, O., Paredes, R. KBoard: Knowledge capture in multimedia collaboration rooms. In *Proc. of Mexican Workshop on Human-Computer Interaction* (MexIHC 2008). *Forthcoming*.

[21] Sánchez, J. A., Strazzulla, D., Paredes R. STRATA: Fostering Effective Collaboration through Multilayered Annotations and Telepointers. In *Proc. of the Eighth Brazilian Symposium on Human Factors in Computer Systems* (IHC 2008, Porto Alegre, Brazil). *Forthcoming*.

[22] Sánchez, J. A., Arzamendi-Pétriz, A., and Valdiviezo, O. 2007. Induced tagging: promoting resource discovery and recommendation in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vancouver, BC, Canada, June 18 - 23, 2007). JCDL '07. ACM, New York, NY (2007), 396-397.

[23] Sánchez, J. A., Quintana, M. G., Razo, A.. Star-fish: Starfields+fisheye visualization and its application to federated digital libraries. *Proceedings of the 3rd Latin American Conference on Human-Computer Interaction* (CLIHC 2007, Nov.), Rio de Janeiro, Brazil.

[24] Silva, J. M. and Favela, J. 2006. Context Aware Retrieval of Health Information on the Web. In *Proc. of the Fourth Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC, 135-146.

[25] Starostenko, O., Chávez-Aragón, A., Burlak, G., Contreras, R. A Novel Star Field Approach for Shape Indexing in CBIR System, *Journal of Engineering Letters, Special Issue on Artificial Intelligence and Computer Science*, International Association of Engineers, Mexico, Vol. 15, Issue 2, (2007), pp. 287-295.

[26] Téllez, E. S., Chávez, E., and Contreras-Castillo, J. SPyRO: Simple Python Remote Objects. In *Proc. of the Fourth Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC (2006), 39-46.

[27] Vázquez, A. R. and Ostróvskaya, Y. A. Analysis of Open Technological Standards for Learning Objects. In *Proceedings of the Fourth Latin American Web Congress*. LA-WEB. IEEE Computer Society, Washington, DC (2006), 105-108.

[28] Villaseñor-Pineda, L., Montes-y-Gómez, M., Pérez-Coutiño, M., and Vaufreydaz, D. A Corpus Balancing Method for Language Model Construction. *Conference on Intelligent Text Processing and Computational Linguistics*, CICLing-2003. Mexico City, February. Lecture Notes in Computer Science, Vol. 2588, Springer, (2003).