

# Toward a better Web Information Retrieval System

Sorin OSTAFIEV

[sorin@ostafiev.com](mailto:sorin@ostafiev.com)

<http://www.ostafiev.com>

October 2003

University "Politehnica" of Bucharest

Department of Computer Science

**Goal:** With the fast growth of the Internet, more and more information is available on the Web and as a result, Web Information Retrieval (IR) has become a fact of life for most Internet users. However, compared with classic information retrieval, web information retrieval systems are faced with totally different datasets (bulk documents, dynamism of the Internet, duplication, heterogeneity, high linkage, ill-formed queries, etc...). It is estimated that nearly 85% users only look at the first screen of the returned results from search engines. 78% users never modify their very first query. So the big challenge to the Web information retrieval is to meet the users information needs given the uniqueness of Web. The performance of an IR system is evaluated along three lines: recall (the percentage of relevant pages that are returned), precision (the percentage of pages retrieved that are relevant) and precision at top 10 result pages. In Web IR, the quality of pages varies widely and thus just being relevant is not enough.

The goal of an efficient Web Information Retrieval system is to return both high-relevance and high-quality (valuable) pages.

There are a few studies in IR literature that use Genetic Algorithms (GAs). The main idea of the proposed research is to exploit the power and speed of GAs to try to solve one or more of the problems and challenges of a Web IR system, such as: ontology document mapping, query representation, matching user requirements (queries) with document representations, and the synthesis of user's profile. A parallel direction of research to be investigated is the possibility to exploit the newly developed concepts of Interactive GAs and KeyGraphs in text document retrieval and Chance Discovery.

# 1. Basic Information Retrieval System

A document based IR system typically consists of three main subsystems:

- **document retrieval and representation**
- **representation of users' requirements (queries)**
- **the algorithms used to match user requirements (queries) with document representations**

A document retrieval engine (also called crawler, spider or robot) collects documents by recursively fetching links from a set of starting pages. Each crawler has different policies with respect to which links are followed, how deeply sites are explored, etc.

A document collection consists of many documents containing information about various subjects or topics of interests. Document contents are transformed into a document representation (either manually or automatically). Document representations are done in a way such that matching these with queries is easy.

Another consideration in document representation is that such a representation should correctly reflect the author's intention. The primary concern in representation is how to select proper index terms. Typically representation proceeds by extracting keywords that are considered as content identifiers and organizing them into a given format.

## **Domain specific text characteristics:**

- Vocabulary set is limited
- Word usage has patterns
- Semantic ambiguities are rare
- Terms and jargon of the domain appear frequently

These characteristics allow us to build ontologies and use machine learning algorithms to extract knowledge.

There are some basic **IR Paradigms**.

- **Probabilistic IR:** Probabilistic retrieval is based on estimating a probability of relevance of a document to the user for the given user query. Typically relevance feedback from a few documents is used to establish the probability of relevance for other documents in the collection. There are three different learning strategies used in probabilistic retrieval. Estimation of probabilities of relevance is done for a set of sample documents, or a set of sample queries and extended to all the documents or queries. Inference networks use a document and query network that capture probabilistic dependencies among the nodes in the network.

- ***Knowledge based IR***: This approach focuses on modeling two areas. First, it tries to model the knowledge of an expert retriever in terms of the expert's domain knowledge, that is, his or her search strategies and feedback heuristics. An example of such an approach is the Unified Medical Language System. Another area that has been modeled is the user of the system. This typically follows the way the librarian develops a client profile. Although knowledge based approaches might be effective in certain domains, it may not be applicable in all domains.
- ***Learning systems based IR***: This approach is based on algorithmic extraction of knowledge or identifying patterns in the data. There are three broad areas within this approach: Symbolic learning, Neural networks and Evolution based algorithms.

Knowledge extraction is a form of text processing that locates a set of relevant items in a natural-language document. Evolutionary algorithms are based on the Darwinian principles of natural selection. These algorithms can be further divided into: GA's, evolutionary strategies, and evolutionary programming. While evolutionary programming utilizes changes at the level of species, the evolutionary strategies exploit changes at individual behavioral level. GAs are based on genetic operators of selection, crossover, and mutation. There are a few studies in IR literature that use GAs. Also, parallel GAs tend to be widely used.

## 2. Functionalities lacking in traditional engines

- ***filtering***: A user looking for some topic on Internet retrieves too much information.
- ***ranking of retrieved documents***: The system provides no qualitative distinction among the documents.
- ***support of relevance feedback***: The user cannot tell his subjective evaluation of the relevance of the document.
- ***personalization***: There is a need of personal systems that serve the specific interest of the users and build users' profiles.
- ***adaptation***: The system should notice when the user changes his/her interests.

### **3. Research directions**

#### **3.1. Ontology document mapping**

The ontological approach used in the treatment of rough information enables a knowledge detection machine to generate domain relevant knowledge organized in XML format. Machine learning algorithms were applied for this purpose by systems like RAPIER, WHISK and HMM, complemented with text processing techniques (like part of speech tagging, text segmentation, noun phrase identification, etc...).

Knowledge extraction module:

- rule based machine learning algorithm
  - RAPIER(Mary Elaine Califf 1997)
  - WHISK(Stephen Soderland 1999)
- Statistical
  - HMM(Hidden Markov Model)

One possible direction of research is to use knowledge extraction based on genetic algorithm to develop ontology based XML knowledge management system that will exploit advanced knowledge discovery techniques and will process text documents residing on Internet, databases, or private document repositories. One possible ontology representation to meet this goal may be RDFS (Resource Description Framework Schema).

#### **3.2. Query representation**

Queries transform the user's information need into a form that correctly represents the user's underlying information requirement and is suitable for the matching process. Query formatting depends on the underlying model of retrieval used: Boolean models, vector space models, probabilistic models, fuzzy retrieval models, models based on artificial intelligence techniques and models including GAs [P. Pathak, M. D. Gordon and W. Fan, 2000].

Previous attempts at using GAs have concentrated on modifying document representations or modifying query representations. One possible direction of research is of applying GAs to adapt various matching functions. Such an adaptation of the matching functions may lead to a better retrieval performance than that obtained by using a single matching function. An overall matching function may be treated as a weighted combination of scores produced by individual matching functions. This overall score may be used to rank and retrieve documents. Weights associated with individual functions may be searched using Genetic Algorithm.

### **3.3. Matching user requirements with document representations**

A matching algorithm matches a user's requests (in terms of queries) with the document representations and retrieves documents that are most likely to be relevant to the user.

A matching algorithm addresses two issues:

- How to decide how well documents match a user's information request. Blair & Maron [1985] showed that it is very difficult for users to predict the exact words or phrases used by authors in desired documents. Hence if a document term does not match search terms then a relevant document may not be retrieved.
- Another issue involved in matching is how to decide the order in which the documents are to be shown to the user. Typically the matching algorithms calculate a matching number for each document and retrieve the documents in the decreasing order of this number.

Both problems may be attacked, presumably, by using genetic algorithms, namely to synthesize new words and/or phrases to perform the search and to select the right order of presenting retrieved documents to the user.

### **3.4. User's profile**

Each web user varies widely in their needs, expectations and knowledge. For best results in web information retrieval, we have to focus on a system with the ability to dynamically adapt to its users. In particular, one possible direction of research is to consider methods for automatically improving the systems recommendation policy on the basis of feedback from the users.

Genetic Algorithms seem to be a good candidate for developing such a method. Interactive Genetic Algorithms (IGA), as the ones presented in [F. C. Hsu, J. S. Chen and P. Chen, 2000] are particularly well suited for such an approach. An IGA is a GA except that fitness function is replaced by human evaluation. A specific feature of the IGA is intended to combine the global search ability of GAs and the evaluation capabilities of humans. Interactive GAs have been used to solve problems that cannot be easily solved by GAs, such as design or art.

### 3.5. Miscellaneous features

Some other challenges in information retrieval that may be investigated during the research are:

#### 3.5.1. Locating the information

- *conceptual search* feature, a revolutionary way of identifying documents by the meaning of the words they contain (ex. search for “ships” and find a document which contains “Titanic”)
- *query disambiguation* feature allows the user to chose an unequivocal meaning for the words in his query (ex. search for “windows” the system will ask to chose between:
  - windows as: open spaces in the wall of a building
  - windows as: an operating system)
- *query increased generality* feature offers the possibility of increasing the power of the conceptual search by finding concepts that are “kind of” the searched concept (ex. search for “pets” and find a document which contains “Tom” which is a “cat” - a kind of “pet”)

#### 3.5.2. Efficient reporting

- **organize the hit list** feature can be used for structuring the hit list in order to speed up the access to the most interesting documents. The result is a tree of documents (a topic map) instead of the simple list.
- **highlighting** feature is used for marking found instances and most relevant paragraphs (with respect to the query).
- **results presentation**
  - simple ranked list
  - tree
    - grouped by domains
    - grouped by derived concepts (in case of ontological subtree search)

#### 3.5.3. Other features

- **document filtering**
- **document categorization**
- **document and knowledge statistics**
- **knowledge processing**

## 4. KeyGraphs and Chance Discovery

### **Chance Discovery** - *discovering chances*

Chance Discovery (CD) means discovering chances - the breaking points in systems, the marketing windows in business, etc. Despite its infancy as a research field, chance discovery has already attracted considerable interest from researchers of various disciplines, including web-related research, finance, and simulation of natural disasters. It involves determining the significance of some piece of information about an event and then using this new knowledge in decision making. The techniques already developed combine data mining methods for finding rare but important events with knowledge management, group-ware and social psychology. May be used for finding information on the Internet, recognizing changes in customer behavior, detecting the first signs of an imminent earthquake, etc.

Chance discovery is mainly motivated from the practical, application-oriented side. However, this field also gives rise to a series of more foundational questions, which seem highly interesting from a methodological view of science.

### **KeyGraph** - *a keyword extraction method*

KeyGraph, originally an algorithm for extracting terms (words or phrases), expresses assertions based on the co-occurrence graph of terms from textual data. The strategy of KeyGraph comes from considering that a document is constructed like a building for expressing new ideas based on traditional concepts.

KeyGraph is a fast method for extracting keywords representing the asserted core idea in a document. KeyGraph composes clusters of terms, based on co-occurrences between terms in a document. Each cluster represents a concept on which the document is based and terms connecting clusters tightly are obtained as author's assertion.

The KeyGraph procedure is a graphical method for data mining originally developed for indexing a document [Y. Ohsawa, N. E. Benson and M. Yachida, 1998] and recently utilized for chance discovery [Y. Ohsawa, 2001, 2002] including discovery of risky active faults of earthquakes [Y. Ohsawa and M. Yachida, 1999]. Like other text and data mining algorithms, KeyGraph identifies relationships between terms and term clusters in a document. In particular, KeyGraph focuses on co-occurrence relationships, but one thing that sets KeyGraph apart is its emphasis on both high and low probability events.

# Glossary

1. **Categorization:** attributing a domain/sub domain to a document
2. **Concept:** an abstract or general idea inferred or derived from specific instances; a class or a category
3. **Conceptual Marked Text:** usually an XML document containing text in which words are identified as concepts or instances; storing links to a specific ontology does this
4. **Conceptual Search:** a method of finding documents containing specific concepts rather than key words
5. **Dictionary:** a list of words stored in machine-readable form, with information given for each word, usually including meaning and POS
6. **Document:** any kind of machine-readable data having a structural or semantic coherence; a source of information
7. **Domain:** a sphere of activity or interest; a field
8. **Free Text:** a piece a text with no topological structure
9. **Genetic Algorithms (GAs):** adaptive methods for solving different problems of searching and optimization. They are based on natural, genetic rules and the process of evolution
10. **Information:** a collection of facts from which conclusions may be drawn; the content of a document
11. **Information Extraction:** a process that takes texts and produces fixed-format, unambiguous data as output
12. **Information Repository:** a collection of documents organized in a manner that facilitates information retrieval
13. **Information Retrieval (IR):** a process that recovers from a collection a subset of documents which are relevant to a query
14. **Instance:** a concrete representation of a concept; an individual object of a certain class
15. **Interactive Genetic Algorithm (IGA):** a GA except that fitness function is replaced by human evaluation
16. **KeyGraph:** a fast method for extracting keywords representing the asserted core idea in a document
17. **Knowledge:** Information represented in machine-understandable format (frames, rules, etc.)
18. **Knowledge Base (KB):** A collection of knowledge, represented using some knowledge representation language
19. **Knowledge Discovery (KD):** nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley 1992); the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data
20. **Knowledge Management:** the collection of processes that govern the creation, dissemination, and utilization of knowledge; newly emerging, interdisciplinary business model dealing with all aspects of knowledge within the context of the firm, including knowledge creation, codification, sharing, and how these activities promote learning and innovation
21. **Knowledge Engineering:** the process of building an expert system, usually consists of three phases: Knowledge Acquisition, Knowledge Elicitation, Knowledge Representation
22. **Machine learning algorithm:** an algorithm that automatically improve its performance with past experiences
23. **Natural Language Processing:** is the engineering of systems that process or analyze written or spoken natural language
24. **Natural Language:** language used in a natural way (the way humans use it when they interact between them)
25. **Ontology:** a description (a formal specification) of the concepts and relationships that exist in specific domain
26. **POS (Part of Speech) tagging:** the process of marking the words in a text specifying the part of speech they represent (nouns, verbs etc.)
27. **RDF/RDFS:** Resource Description Framework / Resource Description Framework Schema; an XML based language here used for describing ontologies
28. **Recommender Systems:** systems developed to adapt web sites gradually to their users. The task of a recommender is to make it easier for the users of a web site to obtain information.
29. **Semantic:** meaning; the sense we give to a word
30. **Structured Document:** a document having a well defined internal structure (ex: a XML document, the result of a database query)
31. **Structured Text:** a piece of text having a rigid topological structure (ex. a table)
32. **Text:** words treated as data by a computer
33. **Unstructured Document:** a document with no relevant internal structure

## References

**„A Comparative Study of Approaches to Chance Discovery“** - Helmut Prendinger and Mitsuru Ishizuka

**„A Comparison of Document Clustering Techniques“** - Michael Steinbach, George Karypis and Vipin Kumar

**„A Survey On Web Information Retrieval Technologies“** - Lan Huang

**„An evaluation of retrieval effectiveness for a full text document-retrieval system“** - David C. Blair and M. E. Maron

**„An overview of web mining“** - Raymond Kosala, Hendrik Blockeel and Frank Neven

**„Discovering Deep Building Blocks for Competent Genetic Algorithms Using Chance Discovery via KeyGraphs“** - David E. Goldberg, Kumara Sastry and Yukio Ohsawa

**„Discovering Emerging Topics from WWW“** - Naohiro Matsumura, Yutaka Matsuo, Yukio Ohsawa and Mitsuru Ishizuka

**„Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation“** - Praveen Pathak, Michael Gordon and Weiguo Fan

**„Interactive Genetic Algorithms For A Travel Itinerary Planning Problem“** - F. C. Hsu, J. S. Chen and P. Chen

**„KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor“** - Yukio Ohsawa, Nels E. Benson and Masahiko Yachida

**„Web Information Retrieval - an Algorithmic Perspective“** - Monika Henzinger

## Preliminary bibliography

- "A Comparative Study of Approaches to Chance Discovery" - Helmut Prendinger and Mitsuru Ishizuka
- "A Comparison of Document Clustering Techniques" - Michael Steinbach, George Karypis and Vipin Kumar
- "A Component-Based Framework For Ontology Evolution" - Michel Klein and Natalya F. Noy
- "A Data Mining Algorithm for Generalized Web Prefetching" - Alexandros Nanopoulos, Dimitrios Katsaros and Yannis Manolopoulos
- "A Declarative Query Interface for Large Semantic Networks" - P. Mork, D. Suci, J. F. Brinkley and C. Rosse
- "A Experiment Report about a Web Information Retrieval System" - Iwao Nagashiro and Dafe ng Cao
- "A Framework for Understanding and Classifying Ontology Applications" - Robert Jasper and Mike Uschold
- "A Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent" - Maria J. Martin-Bautista, Amparo Vila and Henrik Legind Larsen
- "A Generic Approach for Knowledge-Based Information-Site Selection" - Thomas Eiter, Michael Fink, Giuliana Sabbatini and Hans Tompits
- "A Knowledge-based Approach to Information Site Selection" - Thomas Eiter, Michael Fink and Hans Tompits
- "A Maximum Variance Cluster Algorithm" - C.J. Veenman, M.J.T. Reinders and E. Backer
- "A Method for Word Sense Disambiguation of Unrestricted Text" - Rada Mihalcea and Dan I. Moldovan
- "A Model and Architecture for Conceptualized Data Annotations" - Michael Gertz and KaiUwe Sattler
- "A Multi-agent Architecture for Knowledge Management Systems" - Cesar Tacla and Jean-Paul Barthes
- "A Practical Part-of-Speech Tagger" - Doug Cutting, Julian Kupiec, Jan Pedersen and Penelope Sibun
- "A Road map to More Effective Web Personalization - Integrating Domain Knowledge with Web Usage Mining" - Honghua (Kathy) Dai and Bamshad Mobasher
- "A Simple Rule-based Part of Speech Tagger" - Eric Bill
- "A Survey On Web Information Retrieval Technologies" - Lan Huang
- "A Tutorial on Automated Text Categorisation" - Fabrizio Sebastiani
- "A WordNet Based Rule Generalization Engine In Meaning Extracting System" - Joyce Yue Chai and Alan W. Biermann
- "A WordNet-Based Interface to Internet Search Engines" - Dan I. Moldovan and Rada Mihalcea
- "A survey of machine learning approaches to analysis of large corpora" - Xunlei Rose Hu and Eric Atwell
- "Adaptive Sentence Boundary Disambiguation" - David D. Palmer and Marti A. Hearst
- "An Analysis of Ontology Mismatches, Heterogeneity versus Interoperability" - Pepijn R.S. Visser, Dean M. Jones, T.J.M. Bench-Capon and M.J.R. Shave
- "An Automatic Method for Generating Sense Tagged Corpora" - Rada Mihalcea and Dan I. Moldovan
- "An Experiment in Semantic Tagging using Hidden Markov Model Tagging" - Frederique Segond, Anne Schiller, Gregory Grefenstette and Jean-Pierre Chanod
- "An Iterative Approach to Word Sense Disambiguation" - Rada Mihalcea and Dan I. Moldovan
- "An Organization Ontology for Enterprise Modelling" - Mark S. Fox, Mihai Barbuceanu, Michael Gruninger and Jinxin Lin
- "An evaluation of retrieval effectiveness for a full text document-retrieval system" - David C. Blair and M. E. Maron
- "An overview of web mining" - Raymond Kosala, Hendrik Blockeel and Frank Neven
- "Answer Fusion with On-line Ontology Development" - Roxana Girju
- "Applications of a Web Query Language" - Gustavo O. Arocena and Alberto O. Mendelzon
- "Approximate Ontology Translation and its Application to Regional Information Services" - Jun-ichi Akahani, Kaoru Hiramatsu and Kiyoshi Kogure
- "Automated FAQ Answering - Continued Experience with Shallow Language Understanding" - Eriks Sneideris
- "Automated Processing of Structured Online Documents" - Vladimir Kulyukin, Kristian Hammond and Robin Burke
- "Automatic Error Detection in Part of Speech Tagging" - David Elworthy
- "Blockmodeling Techniques for Web Mining" - Gabriella Schoier
- "Book Recommending Using Text Categorization with Extracted Information" - Raymond J. Mooney, Loriene Roy and Paul N. Bennett

"**Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction**" - Mary Elaine Califf and Raymond J. Mooney

"**Building Ontologies - Towards a Unified Methodology**" - Mike Uschold

"**Clustering Algorithms for Spatial Databases - A Survey**" - Erica Kolatch

"**Combining Text and Heuristics for Cost-Sensitive Spam Filtering**" - Jose M. Gomez Hidalgo and Manuel Mana Lopez

"**Combining and relating ontologies - an analysis of problems and solutions**" - Michel Klein

"**Computational complexity of planning and approximate planning in the presence of incompleteness**" - Chitta Baral, Vladik Kreinovich and Raul Trejo

"**Concept Indexing - A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval and Categorization**" - George Karypis and Eui-Hong (Sam) Han

"**Conceptual Clustering of Text Clusters**" - Andreas Hotho and Gerd Stumme

"**Conceptual Models and Architectures for Advanced Information Systems**" - Larry Kerschberg and Doyle J. Weishar

"**Coverage and Competency in Formal Theories - A Commonsense Theory of Memory**" - Andrew S. Gordon and Jerry R. Hobbs

"**Creating, Integrating and Maintaining Local and Global Ontologies**" - Mike Uschold

"**DISCUS - Distributed Innovation and Scalable Collaboration in Uncertain Settings**" - David E. Goldberg, Michael Welge and Xavier Llorca

"**Data Mining on Large Graphs**" - Christopher R. Palmer, Phillip B. Gibbons and Christos Faloutsos

"**Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora**" - Bruno Maximilian Schulze, Ulrich Heid, Helmut Schmid, Anne Schiller, Mats Rooth, Gregory Grefenstette, Jean Gaschler, Annie Zaenen and Simone Teufel

"**Discovering Deep Building Blocks for Competent Genetic Algorithms Using Chance Discovery via KeyGraphs**" - David E. Goldberg, Kumara Sastry and Yukio Ohsawa

"**Discovering Emerging Topics from WWW**" - Naohiro Matsumura, Yutaka Matsuo, Yukio Ohsawa and Mitsuru Ishizuka

"**Discovering Ontologies from Performance Systems**" - Debbie Richards

"**Domain-Specific Knowledge Acquisition from Text**" - Dan Moldovan, Roxana Girju and Vasile Rus

"**Dynamic Information Filtering**" - Patrick Baudisch

"**Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation**" - Praveen Pathak, Michael Gordon and Weiguo Fan

"**Evaluation of Hierarchical Clustering Algorithms for Document Datasets**" - Ying Zhao and George Karypis

"**Evaluation of Recommender Algorithms for an Internet Information Broker based on Simple Association Rules and on the Repeat-Buying Theory**" - Andreas Geyer-Schulz and Michael Hahsler

"**Evolution of Decision Trees**" - Xavier Llorca and Josep M. Garrell

"**Evolutionary Information Retrieval**" - Pawan Lingras

"**Evolutionary computation as a form of organization**" - Alexander Kosorukoff and David Goldberg

"**Explaining Text Clustering Results using Semantic Structures**" - Andreas Hotho, Steffen Staab and Gerd Stumme

"**Exploiting Web Log Mining for Web Cache Enhancement**" - Alexandros Nanopoulos, Dimitrios Katsaros and Yannis Manolopoulos

"**Exploration, Exploitation In Adaptive Recommender Systems**" - Stephan ten Hagen, Maarten van Someren and Vera Hollink

"**Extending Recommender Systems - A Multidimensional Approach**" - Gediminas Adomavicius and Alexander Tuzhilin

"**Finding and characterizing changes in ontologies**" - Michel Klein, Atanas Kiryakov, Danyan Ognyanov and Dieter Fensel

"**Formal Ontology and Information Systems**" - Nicola Guarino

"**Formal Ontology, Conceptual Analysis and Knowledge Representation**" - Nicola Guarino

"**Formalised Elementary Formal Ontology**" - Luc Schneider

"**From Manual to Semiautomatic Semantic Annotation - About Ontology-based Text Annotation Tools**" - M. Erdmann, A. Maedche, H.-P. Schnurr and S. Staab

"Genetic Algorithms for Internet Search - Examining the Sensitivity of Internet Search by Varying the Relevant Components of Genetic Algorithm" - Šešum Vesna and Cvetkoviæ Dragana

"Genetic Algorithms in Information Retrieval" - Vrajitoru Dana

"Hierarchical Taxonomies using Divisive Partitioning" - Daniel Boley

"High Precision Logic Form Transformation" - Vasile Rus

"How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis" - C. Fraley and A. E. Raftery

"Improving the search on the Internet by using WordNet and lexical operators" - Dan I. Moldovan and Rada Mihaicea

"Information Retrieval on the Web - Selected Topics" - Mei Kobayashi and Koichi Takeda

"Information Retrieval" - Hans Mehlh

"Information retrieval on the Web - Tools and algorithmic issues" - Andrei Broder and Monika Henzinger

"Integrating Lexical Knowledge in Learning-Based Text Categorization" - Jose Maria Gomez Hidalgo, Manuel de Buena Rodríguez, Luis Alfonso Ureña Lopez, Maria Teresa Martín Valdivia and Manuel García Vega

"Integrating the WordNet Ontology into a Description Logic System" - Jordi Alvarez

"Interactive Evolutionary Computation - Fusion of the Capabilities of EC Optimization and Human Evaluation" - Hideyuki Takagi

"Interactive Genetic Algorithms For A Travel Itinerary Planning Problem" - Fang-Cheng Hsu, Jiah-Shing Chen and Poren Chen

"KDD for Personalization - PKDD 2001 Tutorial" - Bamshad Mobasher, Bettina Berendt and Myra Spiliopoulou

"KeyGraph - Automatic Indexing by Co-occurrence Graph based on Building Constructive Metaphor" - Yukio Ohsawa, Nels E. Benson and Masahiko Yachida

"Knowing Me, Knowing You - Practical Issues in the Personalisation of Agent Technology" - Stuart Soltysiak and Barry Crabtree

"Knowledge Sources for Word Sense Disambiguation" - Eneko Agirre and David Martínez

"Knowledge-based information retrieval from semi-structured text" - Robin D. Burke, Kristian J. Hammond and Edwin Cooper

"Learning And Generalization In The Creation Of Information Extraction Systems" - Joyce Yue Chai

"Learning Information Extraction Rules for Semi-structured and Free Text" - Stephen Soderland

"Learning Syntactic Rules and Tags with Genetic Algorithms for Information Retrieval and Filtering - An Empirical Basis for Grammatical Rules" - Robert M. Losee

"Learning to Generate Semantic Annotation for Domain Specific Sentences" - Jianming Li, Lei Zhang and Yong Yu

"Linguistic search engine" - Adam Kilgarriff

"Log Mining to Improve the Performance of Site Search" - Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Wei-Ying Ma and Chao-Jun Lu

"Machine Learning for Information Retrieval - Neural Networks, Symbolic Learning, and Genetic Algorithms" - Hsinchun Chen

"Methodological Considerations on Chance Discovery" - Helmut Prendinger and Mitsuru Ishizuka

"Methodology for the Design and Evaluation of Ontologies" - Michael Gruninger and Mark S. Fox

"Mining Association Patterns in Web Usage Data" - Pang-Ning Tan and Vipin Kumar

"Mining On-line Newspaper Web Access Logs" - Paulo Batista and Mário J. Silva

"Mining Patterns from Graph Traversals" - Alexandros Nanopoulos and Yannis Manolopoulos

"Multi-level Rule Discovery from Propositional Knowledge Bases" - Debbie Richards and Usama Malik

"Natural Language Processing in the FAQ Finder System - Results and Prospects" - Robin Burke, Kristian Hammond, Vladimir Kulyukin, Steven Lytinen, Noriko Tomuro and Scott Schoenberg

"Navigation Planning to Guide Concept Understanding in the World Wide Web" - Seiji Yamada and Yukio Ohsawa

"On Knowledgeable Unsupervised Text Mining" - A. Hotho, A. Maedche, S. Staab and V. Zacharias

"On the Insufficiency of Ontologies - Problems in Knowledge Sharing and Alternative Solutions" - Flavio S. Correa da Silva, Wamberto W. Vasconcelos, David S. Robertson, Virginia Brilhante, Ana C. V. de Melo, Marcelo Finger and Jaume Agustí

"On the integration of technologies for capturing and navigating knowledge with ontology-driven services" - Yannis Kalfoglou, John Domingue, Leslie Carr, Enrico Motta, Maria Vargas-Vera and Simon Buckingham Shum

"**On-To-Knowledge - Semantic Web Enabled Knowledge Management**" - Dieter Fensel, Frank van Harmelen, Ying Ding, Michel Klein, Hans Akkermans, Jeen Broekstra, Arjohn Kampman, Jos van der Meer, York Sure, Rudi Studer, Uwe Krohn, John Davies, Robert Engels, Victor Iosif, Atanas Kiryakov, Thorsten Lau, Ulrich Reimer and Ian Horrocks

"**On-To-Knowledge - Technical Fact Sheet for the OTK Tool Suite**" - On-To-Knowledge Consortium

"**On-To-Knowledge Methodology - Baseline Version**" - Hans-Peter Schnurr, York Sure, Rudi Studer and Hans Akkermans

"**On-To-Knowledge Methodology - Employed and Evaluated Version**" - York Sure and Rudi Studer

"**On-To-Knowledge Methodology - Final Version**" - York Sure and Rudi Studer

"**OntoWeb - Ontology-based information exchange for knowledge management and electronic commerce**" - Alain Léger, Yannick Bouillon, Philippe Ecoublet, Martin Bryan, Rose Dieng, Andreas Persidis, York Sure, Asuncion Gomez-Perez and Mariano Fernández López

"**OntoWeb - a Semantic Web Community Portal**" - P. Spyns, D. Oberle, R. Volz, J. Zheng, M. Jarrar, Y. Sure, R. Studer and R. Meersman

"**Ontological Structures for Knowledge Sharing**" - M. J. R. Shave

"**Ontologies - Principles, Methods and Applications**" - Mike Uschold and Michael Gruninger

"**Ontologies and Databases - More than a Fleeting Resemblance**" - Robert Meersman

"**Ontologies to Support Process Integration in Enterprise Engineering**" - Michael Gruninger, Katy Atefi and Mark S. Fox

"**Ontology Engineering for Active Catalog**" - Jihie Kim, S. Ringo Ling and Peter Will

"**Ontology Evolution - Not the Same as Schema Evolution**" - Natalya F. Noy and Michel Klein

"**Ontology Reuse and Application**" - Mike Uschold, Mike Healy, Keith Williamson, Peter Clark and Steven Woods

"**Ontology versioning and change detection on the Web**" - Michel Klein, Dieter Fensel, Atanas Kiryakov and Damyan Ognyanov

"**Ontology versioning on the Semantic Web**" - Michel Klein and Dieter Fensel

"**Ontology-based Content Management in a Virtual Organization**" - Peter Mika, Victor Iosif, York Sure and Hans Akkermans

"**Ontology-based Text Clustering**" - A. Hotho and S. Staab A. Maedche

"**Ontology-based Text Document Clustering**" - Andreas Hotho, Alexander Maedche and Steffen Staab

"**Overview of Data Mining and Knowledge Discovery in Databases (KDD)**" - William H. Hsu

"**Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing**" - A. Gómez Pérez, M. Gruninger, H. Stuckenschmidt and M. Uschold

"**Question Answering from Frequently-Asked Question Files - Experiences with the FAQ Finder System**" - Robin D. Burke, Kristian J. Hammond, Vladimir A. Kulyukin, Steven L. Lytinen, Noriko Tomuro and Scott Schoenberg

"**Ranked Retrieval with Semantic Networks and Vector Spaces**" - Vladimir A. Kulyukin and Amber Settle

"**Reasoning about Evolving Nonmonotonic Knowledge Bases (INFSYS Research Report)**" - Thomas Eiter, Giuliana Sabbatini, Michael Fink and Hans Tompits

"**Reasoning about Evolving Nonmonotonic Knowledge Bases**" - Thomas Eiter, Michael Fink, Giuliana Sabbatini and Hans Tompits

"**Scalability and Knowledge Reusability in Ontology Modeling**" - Mustafa Jarrar and Robert Meersman

"**Scaling Question Answering to the Web**" - Cody C. T. Kwok, Oren Etzioni and Daniel S. Weld

"**Seed Ontologies - growing digital libraries as distributed, intelligent systems**" - Peter Weinstein and Gene Alloway

"**Semantic Indexing using WordNet Senses**" - Rada Mihalcea and Dan I. Moldovan

"**Semantic Networks - Visualizations of Knowledge**" - Roger Hartley and John Barnden

"**Semantic Networks for Knowledge Representation in an IE**" - Stephen Peters

"**Semantic Networks, Case and Logical Form**" - Robert Wilensky

"**Semantic Networks**" - D. Dankel

"**Some Experiences on Large Scale Web Mining**" - Masaru Kitsuregawa, Iko Pramudiono, Yusuke Ohura and Masashi Toyoda

"**Specification and Underspecification in Lexical Semantic Processing for Information Extraction**" - Paul Buitelaar

"**Structured Representations**" - Melissa Libertus, Sepideh Sadaghiani and Klaus Tichacek

"**Supporting evolving ontologies on the Internet**" - Michel Klein

"Survey Paper - The Development of a Research Network Information System Specification Method" - Aldo de Moor

"Survey of Clustering Data Mining Techniques" - Pavel Berkhin

"Text Clustering Based on Background Knowledge" - Andreas Hotho, Steffen Staab and Gerd Stumme

"Text Clustering Based on Good Aggregations" - Andreas Hotho, Alexander Maedche and Steffen Staab

"The Anatomy of a Large-Scale Hypertextual Web Search Engine" - Sergey Brin and Lawrence Page

"The Ontolingua Server - a Tool for Collaborative Ontology Construction" - Adam Farquhar, Richard Fikes and James Rice

"The PKDD Discovery Challenges on Thrombosis Data" - Petr Berka

"The Role of Identity Conditions in Ontology Design" - Nicola Guarino

"The Use of Lexical Semantics in Information Extraction" - Joyce Yue Chai and Alan W. Biermann

"The Xerox Part-of-Speech Tagger" - Doug Cutting and Jan Pedersen

"Thesauri and Semantic Networks" - Michael Lee

"Three Approaches for Knowledge Sharing - A Comparative Analysis" - Mike Uschold, Rob Jasper and Peter Clark

"Toward Principles for the Design of Ontologies Used for Knowledge Sharing" - Thomas R. Gruber

"Towards Semantic Web Mining" - Bettina Berendt, Andreas Hotho and Gerd Stumme

"Understanding top-level ontological distinctions" - Aldo Gangemi, Nicola Guarino, Claudio Masolo and Alessandro Oltramari

"Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging" - Eric Brill

"Usage Mining for and on the Semantic Web" - Gerd Stumme, Andreas Hotho and Bettina Berendt

"Using Content-Based Filtering for Recommendation" - Robin van Meteren and Maarten van Someren

"Using Genetic Algorithms to Get the Classes and their Number in a Partitional Cluster Analysis of Large Data Sets" - J. A. Lozano and P. Larranaga

"Using Natural Language Interfaces" - William C. Ogden and Philip Bernick

"Using Ontologies for Knowledge Management - An Information Systems Perspective" - Igor Jurisica, John Mylopoulos and Eric Yu

"Using Semantic Networks for Knowledge Representation in an Intelligent Environment" - Stephen Peters and Howard E. Shrobe

"WEB community mining and WEB log mining - Commodity Cluster based Execution" - Masaru Kitsuregawa, Masashi Toyoda and Iko Pramudiono

"Web Information Retrieval - an Algorithmic Perspective" - Monika Henzinger

"Web Information Retrieval" - Monika Henzinger

"Web Mining - Information and Pattern Discovery on the World Wide Web" - R. Cooley, B. Mobasher and J. Srivastava

"Web Mining Research - A Survey" - Raymond Kosala and Hendrik Blockeel

"Web Mining and Knowledge Discovery of Usage Patterns" - Yan Wang

"Web Mining in Soft Computing Framework - Relevance, State of the Art and Future Directions" - Sankar K. Pal, Varun Talwar and Pabitra Mitra

"What is a word? What is a sentence? - Problems of Tokenization" - Gregory Grefenstette and Pasi Tapanainen

"Why Ontologies are not Enough for Knowledge Sharing" - Flavio S. Correa da Silva, Wamberto Weber Vasconcelos, Jaume Agusti, David Robertson and Ana Cristina V. de Melo

"Word Sense Disambiguation And Its Application To Internet Search" - Rada Mihalcea

"Word Sense Disambiguation Using Conceptual Density" - Eneko Agirre and German Rigau

"Word Sense Disambiguation based on Semantic Density" - Rada Mihalcea and Dan I. Moldovan