2011 International Nuclear Atlantic Conference - INAC 2011
Belo Horizonte, MG, Brazil, October 24-28, 2011
ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN
ISBN: **978-85-99141-04-5**

NUCLEAR EXPERT WEB MINING SYSTEM:

MONITORING AND ANALYSIS OF NUCLEAR ACCEPTANCE BY INFORMATION RETRIEVAL AND OPINION EXTRACTION ON THE INTERNET

Thiago Reis[1], Antonio C. O. Barroso[2] and Kengo Imakuma[3]

Instituto de Pesquisas Energéticas e Nucleares, IPEN - CNEN/SP
Av. Professor Lineu Prestes, 2242
05508-000 São Paulo, SP
[1] thiagoreis@usp.br
[2] barroso@ipen.br
[3] kimakuma@ipen.br

ABSTRACT

This paper presents a research initiative that aims to collect nuclear related information and to analyze opinionated texts by mining the hypertextual data environment and social networks web sites on the Internet. Different from previous approaches that employed traditional statistical techniques, it is being proposed a novel Web Mining approach, built using the concept of Expert Systems, for massive and autonomous data collection and analysis. The initial step has been accomplished, resulting in a framework design that is able to gradually encompass a set of evolving techniques, methods, and theories in such a way that this work will build a platform upon which new researches can be performed more easily by just substituting modules or plugging in new ones. Upon completion it is expected that this research will contribute to the understanding of the population views on nuclear technology and its acceptance.

1. INTRODUCTION

Although nuclear energy is currently a well-established electric power production method in many countries, the "nuclear theme" often raises major concerns in the population about its associated risks. Public opinion about nuclear power and its related topics can influence, positively or negatively, institutions and governments views and decisions with respect to nuclear power and applications development in the present days and in the future.

Recently, the Fukushima Dai-ichi nuclear power plant accident evidenced this fact, where many people over the world expressed their views and opinions about the nuclear energy. Apart from traditional media, these manifestations were especially observed and massively registered in web pages, blogs, forums, and social networking web sites on the Internet, creating a valuable information repository about public opinion. As doubtless, the Internet continues to evidence itself as a very important channel for mass communication and social interaction; it consequently becomes a unique, real time updated, data base for collecting information about nuclear acceptance.

However, due to the characteristics of the Internet information basis (hugely sized, heterogeneous, noisy, multilinguistic and dynamic), finding and collecting relevant

information, and especially monitoring opinion sources, in a large-scale manner, are nontrivial tasks. This presents both great challenges and opportunities for mining and discovery information and knowledge from its unstructured and noisy data. Efforts and techniques directed to the above mentioned problems are usually referred as pertaining to a new research field called Web Mining.

This paper presents a research initiative that aims to collect nuclear related information and to analyze opinionated texts by mining the hypertextual data environment and social networks web sites on the Internet. Instead of using traditional statistical techniques, it is proposed a novel Web Mining approach, built around the concept of Expert Systems, for massive and autonomous data collection and analysis.

But, to accomplish these goals it is necessary to combine a set of techniques, methods, and tools from subfields of Web Mining, Artificial Intelligence, and Nuclear Knowledge Management, as well as a substantial effort regarding to the computational implementation of these algorithms. Furthermore, the subfield of Opinion Mining is very recently, so no mature methods have yet been developed for solving many of its problems and, on the other hand, these methods are essential for the nuclear opinion extraction tasks of this work.

Thus, as consequence of its inherent complexity, this work requires a gradual approach to: aggregate the information needs of the nuclear sector; implement the needed techniques and tools of the above mentioned fields, and develop the computational system.

So, firstly we have designed a framework capable of supporting the identified needs, goals, and tackling the foreseen difficulties. This framework represents the initial methodological approach, defining the set of methods and procedures to be used in this research, as well as the integration interfaces between them. It is also used as the basis for the computational system architecture. Next, the computational implementation of algorithms and software programming will be done, followed by the evaluation of the system accuracy in terms of the information retrieval and opinion extraction tasks. Finally, the data collection is executed by the system and the analysis of nuclear “public opinion” is done using statistics.

This framework and computational system are called the Nuclear Expert Web Mining System, which supports and integrates Web Crawling and Machine Learning algorithms, to accomplish information retrieval tasks, and Opinion Mining and Natural Language Processing methods, to accomplish opinion extraction tasks.

2. METHODOLOGY

The four main Web Mining tasks, the framework was designed for, are described as follows:

(1) Web crawling task, regarding to search and collect webpages by browsing the hypertextual web graph;
(2) Webpage topic identification to decide whether a webpage contains information related to the nuclear domain;
(3) Webpage opinion identification to the decide whether a webpage text presents an opinionated content;

(4) Webpage opinion polarity task, regarding to the identification if the opinionated text is positive, negative or neutral.

Also, this framework supports seven specific properties to cope with needs, goals, and difficulties previously discussed. They can be described as the capabilities to be:

a. autonomous, taking its actions independently of human intervention;
b. recurrent, running either continuously or successively;
c. scalable, able to continue working efficiently even with the significant increase in the amount of collected information that is foreseen in the next couple of years;
d. adaptive, able to adapt to the environment of information;
e. expert, within a defined nuclear lexical knowledge domain;
f. persistent, able to store the collected information and metadata;
g. extensible:
    - allowing changes in the created models; and
    - allowing addition of new methods.

There are several Web Mining methods to perform the listed tasks and they can be integrated in a variety of combinations. A choice, based on a good combination of performance features and implementation feasibility, capable of fulfilling the above defined properties, was made for the framework.

Considering the existing web crawling methods for information retrieval, such as: Breadth-First-Search [1], Fish-Search [2], Shark-Search [3]; an adaptation of the renowned, state-of-the-art focused web crawling algorithm called InfoSpider [4][5] was chosen as the basis method to accomplish the tasks (1) and (2). The OpinionObserver [6][7] algorithm is used as the basis method to accomplish the tasks (3) and (4).

The set of methods based on InfoSpider and OpinionObserver can attend to the properties (a), (b), (c) and (d). For property (e) an Expert System will be used, due the need of a deep, structured, and extensible lexical knowledge of nuclear domain, gathered from Nuclear Experts, for the computational inference and autonomous evaluation of the information relevance and opinion identification. Compliance with (f) was achieved by a proper design of the database to store the collected webpages and the process metadata. The interactions of the framework components are designed so that a change or improvement in one component does not affect the operation of the others, attending to the property (g). This way, advances in related research fields and new methods being developed can be integrated to the framework in order to improve the results of information retrieval and opinion extraction.

2.1. Methodological Framework

The basic data unit in this framework is the webpage. All the Web Mining tasks are performed at the webpage-level. As the webpages and their hyperlinks composes the web graph, the main data structure in this framework is a graph, where each node is a webpage and each edge is a hyperlink that connects a webpage to another webpage.

In this framework there are three major elements that use as feedstock the webpage data unit and manipulate the graph data structure: (1) the Internet, (2) the Nuclear Experts, and (3) the Expert System.

First, the Internet basically provides the system with webpages for the Web Mining process. These hypertextual data contain information and opinions about the nuclear sector.

Next, a panel of Nuclear Experts will have the role of "tutors" of the Expert System, being responsible for three tasks: (1) to provide the lexical knowledge of the nuclear domain to the system, (2) to teach the system through feedbacks about the collected webpages relevance and opinion identification, and (3) to perform the evaluation of the system performance regarding to the information retrieval and opinion extraction tasks.

Last, the Expert System is responsible for the integration of the Web Mining methods and for the execution of the four previously defined Web Mining tasks.

The Expert System relies on four core components: (1) the Knowledge Base, (2) the Inference Engine, (3) the Web Crawler, and (4) the Database.

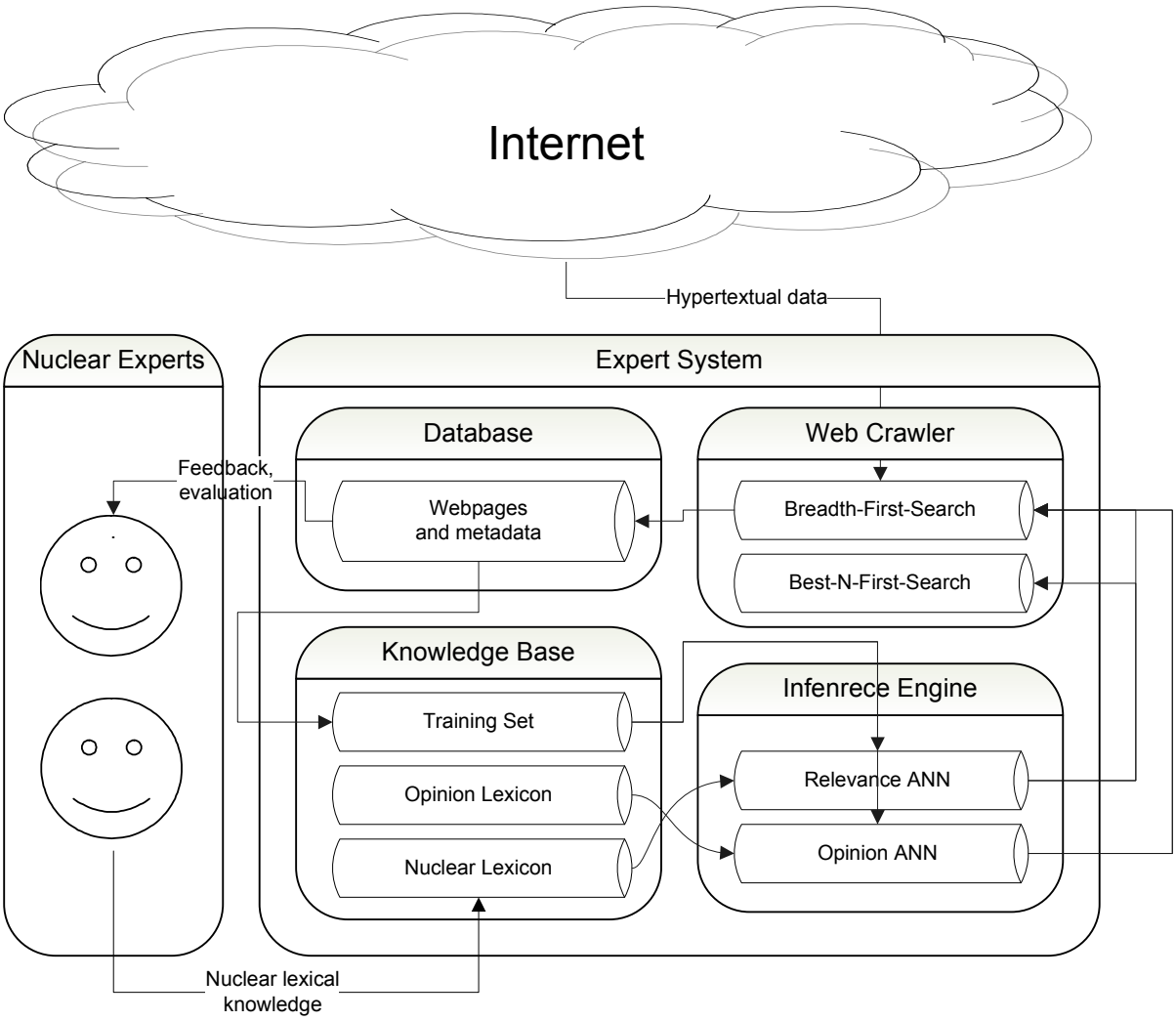

Figure 1. Methodological framework.

The Knowledge Base is the main component of the Expert System because it contains the all a priori knowledge about the nuclear domain gathered from the Nuclear Experts and that serves as the heuristics used by the Inference Engine. The Knowledge Base is composed of three parts: (1) the Training Set, (2) the Nuclear Lexicon, and (3) the Opinion Lexicon.

The Training Set is a set of webpages previously collected and manually classified by the Nuclear Experts regarding to their relevance to nuclear domain and to their opinion polarity. The Nuclear Lexicon is a set of words related to the nuclear domain and represents the knowledge of the Nuclear Experts regarding to which words best discriminate the relevant webpages from the not relevant. It is modeled through interviews with Nuclear Experts, according to the process of knowledge acquisition [8][9], as follows:

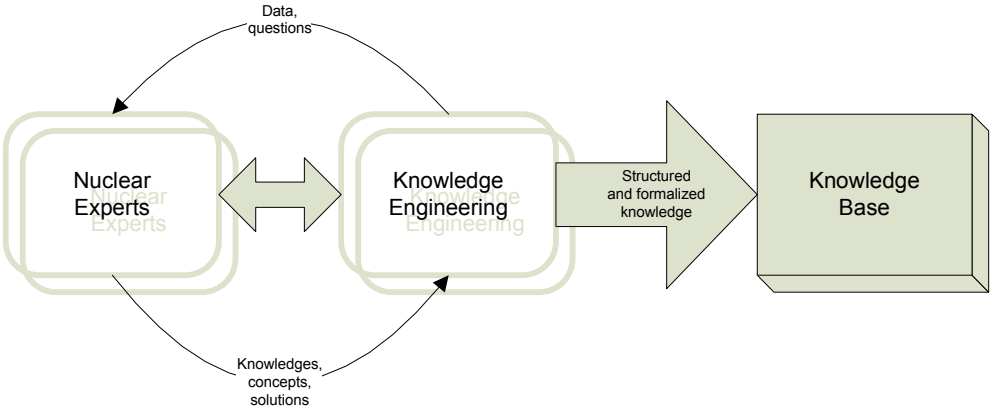

Figure 2. Knowledge acquisition process to build the Nuclear Lexicon.

The Opinion Lexicon consists of adjective words classified according to the prediction of their polarity (semantically positive or negative) and is obtained from the lexicon used by OpinionObserver.

The Knowledge Base is designed to accept new webpages into the Training Set and also new words into the Nuclear and Opinion Lexicons, attending to the property (g).

In addition to a Knowledge Base, an Expert System must have an Inference Engine that simulates the judgmental ability of an expert in a particular task by using the Knowledge Base. The tasks (2), (3) and (4) fit in this case. The Inference Engine performs the task (2) by computing the webpage Relevance Score, which represents the estimated webpage relevance regarding to the nuclear domain. The tasks (3) and (4) are performed by computing the webpage Opinion Score that identifies if the webpage contains opinionated text and what is its opinion polarity. The Training Set is used by the Inference Engine to run the supervised learning algorithms of the machine learning methods. The Nuclear and Opinion Lexicons are used by the Inference Engine to estimate and compute the Relevance and Opinion Scores. A third score, called Similarity Score, is computed in the task (1) and is used to perform the reinforcement learning of the machine learning methods.

The machine learning method to be initially implemented into the Inference Engine in this framework to compute the scores is an artificial neural network (ANN). However, as this framework is built under the webpage data unit conception, others machine learning methods (bayesian classifier, support vector machines) can be implemented and added into the Inference Engine to compute the Relevance and Opinion Scores at the webpage-level and to perform these tasks.

The ANN that computes the Relevance Score is called Relevance ANN (RANN) and is an adaptation of the model used by InfoSpider. The ANN that computes the Opinion Score is called Opinion ANN (OANN). Both ANNs are a single-layer feedforward perceptron and perform their tasks through the analysis of the words within the webpage.

The Web Crawler is the component responsible for collecting information on the Internet by "browsing" the web graph formed by the webpages and their hyperlinks. To this end, the Web Crawler implements a search algorithm that defines the way in which the webpages will be collected. Initially in this framework two search algorithms will be implemented: (1) Breadth-First Search (BFS) and (2) Best-N-First Search (BNFS). However, others graph search algorithms can be implemented to manipulate the graph data structure of the framework.

The BFS performs an exhaustive search, navigating across all hyperlinks and collecting all webpages regardless of their content. The BNFS performs a heuristic search, selecting which hyperlink to navigate and collecting only webpages that possibly have relevant information to the nuclear domain. The RANN output Relevance Score is used by the BNFS to select which next hyperlink is to be followed and therefore which target webpage is to be collected. Both algorithms are based on the Problem Solving by Graph Search algorithms from Artificial Intelligence.

The Database used is a relational database that stores the collected webpages and the metadata generated in the process (graph structure of the collected webpages, webpages Relevance, Opinion and Similarity Scores, ANNs synaptic weights, etc). The collected webpages can be evaluated about their relevance to nuclear domain and opinion polarity by the Nuclear Experts. If so, these webpages are then integrated to the Knowledge Base into the Training Set.

As stated in the property (g) of the framework, the components of the Expert System can be changed and improved independently in order to follow the advances in related research fields and to refine the results of information retrieval and opinion extraction tasks.

2.2. Mining Algorithm

In the framework mining algorithm, the Nuclear and Opinion Lexicons together with the webpages are the inputs of the Inference Engine. Specifically for the ANN method, each RANN's input node is associated with a word contained in the Nuclear Lexicon. The RANN output is the score representing the estimated relevance of the target webpage and is computed through the activation function tanh (hyperbolic tangent), so its output is a real number in the interval from -1 to 1. The tanh function is adopted because it can model both a positive match as a negative match between input words and the estimated relevance of a webpage [4][5].

The RANN input is computed as follows: First, the webpage hyperlinks are extracted. Next, for each webpage outgoing hyperlink, each input node of the RANN is computed by counting the words in the webpage that match the existing words in the Nuclear Lexicon, and each word in the Nuclear Lexicon, in turn, corresponds to a RANN input node. This count is weighted by weights that decay as the distance from the word to the webpage outgoing hyperlink increases, within a window of size *p*. Thus, for each hyperlink *l* and for each word *k*, the RANN receives as input [4][5]:

$$\mathrm{in}_{k,l} = \sum_{i:\mathrm{dist}(k_i,l)\leq\rho} \frac{1}{\mathrm{dist}(k_i, l)} \tag{1}$$

where *ki* is the *i*-th occurrence of the word *k* in the webpage *D* and *dist(ki, l)* is the count of hyperlinks between *ki* and *l* (including *l* up to a maximum distance of *p* hyperlinks).

The Relevance Score is the first estimation of the webpage relevance. After collecting the target webpage, the Similarity Score is computed as a second estimate of its relevance, using the words contained therein and which were previously unknown, by computing the following Cosine Similarity function:

$$\mathrm{sim}(q,p) = \frac{\sum_{k\epsilon q\cap p} f_{kq} f_{kp}}{\sqrt{\left(\sum_{k\epsilon p} f_{kp}^2\right)\left(\sum_{k\epsilon q} f_{kq}^2\right)}} \tag{2}$$

where *q* is the word, *p* is the webpage and *fkd* is the frequency of word *k* in *d*. The Similarity Score result is used as a reinforcement signal from the environment and as expected value of RANN output to calculate the RANN error and be able to run the online backpropagation learning algorithm, thus constituting a reinforcement learning method [4][5].

The OANN has two input nodes. The first node is associated with the positive opinion words and the second node is associated with negative opinion words both contained in the Opinion Lexicon [6][7]. The OANN's output is the score representing the opinion polarity of the text within the target webpage and is computed likewise the RANN. However, in contrast to the RANN, a negative output represents a negative opinion and a positive output represents a positive opinion. An output near zero means that the webpage do not contains opinionated text.

The OANN input is computed as follows [6][7]: First, each word in the webpage is tagged according to their part-of-speech class by using a POS Tagger. Next, the words tagged as adjectives are grouped into a positive set and into a negative set according to their predicted polarity in the Opinion Lexicon and the word counting in each set is computed. If there is a negation word such “no”, “not”, “yet” appearing in a window of N words around the adjective, its polarity is set to its opposite sense. Finally, the word counting of each set is used

as input into OANN and the opinion identification and polarity is estimated and represented by the OANN output score.

The main steps of the framework mining algorithm are:

1 Initialize Web Crawler by using a seed hyperlinks set
  1.1 Collect each webpage referenced by each hyperlink of the seed hyperlinks set
  1.2 Compute the Opinion Score for each collected webpage by using the machine learning algorithm of the Inference Engine
  1.3 Store each collected webpage and its metadata in the Database
  1.4 Extract all the hyperlinks within each of the collected webpages
  1.5 Compute the Relevance Score for each hyperlink by using the machine learning algorithm of the Inference Engine
  1.6 Enqueue each hyperlink in the Web Crawler according to the its Relevance Score and the search algorithm used

2 Recursively run Web Crawler algorithm
  2.1 Dequeue the hyperlinks from Web Crawler queue according to the search algorithm
  2.2 Collect each webpage referenced by each hyperlink of the dequeued hyperlinks set
  2.3 Compute the Similarity Score for each collected webpage by using the Cosine Similarity function
  2.4 Run the reinforcement learning for the machine learning algorithm by using the Similarity Score
  2.5 Compute the Opinion Score for each collected webpage by using the machine learning algorithm of the Inference Engine
  2.6 Store each collected webpage and its metadata in the Database
  2.7 Extract all the hyperlinks within each of the collected webpages
  2.8 Compute the Relevance Score for each extracted hyperlink by using the machine learning algorithm of the Inference Engine
  2.9 Enqueue each extracted hyperlink in the Web Crawler queue according to its Relevance Score and the search algorithm

3 Run the Inference Engine supervised learning
  3.1 Select Nuclear Experts new feedbacked webpages from the Training Set
  3.2 Run the supervised learning for the Relevance Score
  3.3 Run the supervised learning for the Opinion Score

In the first step of the mining algorithm, the Web Crawler is initialized by using a seed hyperlinks set. The webpages that the hyperlinks targets to are collected, stored in the Database, and their Opinion Scores are computed by the Inference Engine. Then, the hyperlinks within the collected webpages are extracted, their Relevance Scores are computed by the Inference Engine, and they are enqueued in the Web Crawler according to the their Relevance Score and the search algorithm used.

In the second step of the mining algorithm, the Web Crawler recursively performs the following steps: first, dequeue the hyperlinks from Web Crawler queue according to the search algorithm. The webpages that the hyperlinks targets to are collected, their Similarity Scores are computed, and the ANN reinforcement learning is performed by using the Similarity Score. Next, the Opinion Scores are computed by the Inference Engine and the

webpages are stored in the Database. Finally, the hyperlinks within the collected webpages are extracted, their Relevance Scores are computed by the Inference Engine, and they are enqueued in the Web Crawler according to the their Relevance Score and the search algorithm used.

In the third step of the mining algorithm, the Inference Engine supervised learning is performed by using the Nuclear Experts new feedbacked webpages from the Training Set.

2.3. Performance Evaluation

The four defined tasks can be understood as classification tasks, therefore, to evaluate the effectiveness of system, are used two classification metrics: Recall and Precision. These metrics are derived from the counting of true positives, true negatives, false positives and false negatives from the classifications results.

To perform these counting, the Training Set is build by crawling and manually classifying a significant number of webpages. Then, the system uses the Training Set as a simulation environment, running inside it. A cutoff in the Relevance Scores and Opinion Scores is defined to classify the collected webpages in the following classes: Relevant, Not Relevant, Positive Opinion, Negative Opinion, and Neutral.

Finally, the metrics are analyzed and the system is executed until the Inference Engine is trained and the metrics are satisfactory.

2.4. Stages of Development

The Nuclear Expert Web Mining System project will be developed in three stages. The first stage, which is the subject of this paper, is the framework design and is already done, as presented.

The second stage is currently started, regarding to the software programming and algorithms implementation together with the Nuclear Lexicon modeling. The software is developed using the Java Programming Language, thus keeping its operational system portability and object-oriented building. A preliminary version of the system is expected to be released in the next months.

The third step is supposed to start as soon as a stable version of the software is released. In this step, the system experimental evaluation is performed and a long crawl will be run to collect a significant amount of data. Then, the nuclear acceptance analysis is done by the analysis of the collected data.

3. CONCLUSIONS

The Internet represents a new and measurable source of information, becoming an important channel for mass communication and social interaction and so an important tool for collecting information about nuclear acceptance. However, finding and collecting its information, and especially monitoring opinion sources, in an autonomous and large-scale manner, are nontrivial tasks.

The purpose of this paper is to present our research initiative that aims to collect nuclear domain related information and to analyze opinionated texts on the Internet and its first achievement regarding to the design of a methodological framework able to gradually encompass the information needs of the nuclear sector, the developments and advances of theories and methods of the related research fields, and the project of the computational system.

Thus, this is an initial research result concerning to the study and organization of the existing methods that can contribute to the understanding of the population views on nuclear sector by analyzing how the Internet can be used as a nuclear acceptance data source and how these methods can be integrated to form a consistent methodological approach. For next steps, we will perform the computational implementation and performance evaluation of these methods followed by the system data collection and nuclear acceptance statistical analysis.

The direct benefit of this research is to provide a means to collect, analyze and monitoring public opinion in a massive and autonomous manner by mining the Internet hypertextual data. Indirectly, this research expects to contribute to nuclear sector by providing a computational system able to search the Internet for nuclear related information and also to the development of related research fields as a singular research at the intersection of Web Mining, Artificial Intelligence, and Nuclear Acceptance fields.

REFERENCES

1. B. Pinkerton, “Finding what people want: Experiences with the WebCrawler,” *Proc. of the 1st International World Wide Web Conference*, Geneva, May 25-26-27 (1994).
2. P. De Bra, and R. Post, “Information retrieval in the World Wide Web: Making clientbased searching feasible,” *Proc. of the 1st International World Wide Web Conference*, Geneva, May 25-26-27 (1994).
3. M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur, “The shark-search algorithm - An application: Tailored Web site mapping,” *Proc. of the 1st International World Wide Web Conference*, Geneva, May 25-26-27 (1994).
4. F. Menczer, “Complementing search engines with online Web mining agents,” *Decision Support Systems*, **35**, pp. 195-212 (2003).
5. F. Menczer, G. Pant, M. Ruiz, and P. Srinivasan, “Evaluating topic-driven Web crawlers,” *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, pp. 241-249 (2001).
6. M. Hu, and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Aug 22-25 (2004).
7. B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proc. of the 14th International World Wide Web Conference*, Chiba, May 10-14 (2005).
8. P. Jackson, *Introduction to Expert Systems*, Addison Wesley, pp. 4-6 (1998).
9. R. Akerkar, P. Sajja, *Knowledge-Based Systems*, Jones & Bartlett Publishers, pp. 60-68 (2009).