# Bayesian Semantics Incorporation to Web Content for Natural Language Information Retrieval

**Manolis Maragoudakis and Nikos Fakotakis**
Intelligent Systems Group
University of Patras
Rion 26500, Patras, Greece
{mmarag,fakotaki}@wcl.ee.upatras.gr

**Abstract**

For the present work, we endeavor with the important aspect of information retrieval of Web content using natural language queries. Currently, markup languages and formalisms do not fully provide mechanisms for effective and accurate analysis of Web content but rather provide means for describing the content in a more human-centric approach. As a result, natural language queries cannot be handled by the Internet search engines. Other approaches use grammar markup labels that attempt to fully match an unforeseen query. For the purposes of this paper, we introduce the theoretical and implementation issues of a novel, statistical framework that can cope with Web content analysis and information retrieval using natural language. The framework is based on Bayesian networks, a tool for knowledge representation and reasoning under conditions of uncertainty. The Web page designer provides the lexical items that contain useful information and labels the corresponding semantic interpretation, from a pre-defined set of domain categories. This knowledge is used for learning the structure and the parameters of a Bayesian network. At the time a user's query is encountered, the network is used in order to return pages that contain the most related semantic content to the user's query.

## 1. Introduction

Web content has been considered for human utilization through the plethora of computers connected to it. As technology evolves, Internet connectivity incorporates new intelligent devices such as mobile phones, robots, and Personal Digital Assistants. Due to limitations of the physical dimensions of such devices, it is reasonable to state that interaction with the Internet should become more human-centric. The current expansion in Web content has not been accompanied by a development of mechanisms that will provide intelligent agents the means for effective and accurate analysis of that content. Under this perspective, natural language (NL) interaction emerges as an effective way of retrieving information. Nevertheless, contemporary search and content analysis engines do not reply to NL queries with precise responses. Probably the most plausible reason is that current markup languages do not provide the means to search engines to utilize NL interactivity. Instead, they help search engines provide links to content that is closely related to the keywords found in a query. Moreover, current engines incorporate a context-specific format in order to overcome the exact keyword matches, meaning that they require users to learn a new way of interaction that is a far cry from NL interaction. Due to the fact that the Internet is a conglomeration of information sources, it is almost impractical for intelligent agents to analyze this content at the time a user seeks for information. We suggest a statistical skeleton that lets Internet agents discover accurate, concise content and respond to NL queries. Its backbone consists of Bayesian networks that provide a statistical, yet semantically-oriented representation of information content. They reflect information in Web pages by anticipating the semantic interpretation of a user query to retrieve related content.

### 1.1 Background

Hypertext Markup Language (HTML) was the initial language for document presentation on the Web and is still the most widely-accepted language on the Internet. HTML inherently lacks the semantics to permit agents to comprehend the knowledge on the Internet. Researchers have applied Natural Language Processing (NLP) techniques to understand text content, but with partial success (Soderland, 1997). Recent attempts to augment the Web content with semantic information by embedding special fields, called *tags*, has led to the development of the so-called *Semantic Web*. The Semantic Web's (Fensel, 2000) realization is underway with the development of AI-inspired content description markup languages. Extensible Markup Language (XML) is the foundation for all recent efforts to create the Semantic Web. XML uses a prose description to imply meaning in documents. The necessity for uniform semantics that all search engines can understand led to the development of the Resource Description Framework (RDF). RDF uses metadata to unambiguously describe Web content. Simple HTML Ontological Extensions (SHOE) (Heflin, Hendler and Luke, 1999) is also based on the frame system. SHOE lets authors use Horn clause logic to annotate content. The Ontology Inference Layer (OIL) lets authors use descriptions to assert different kinds of definitions. The DARPA Agent Markup Language and DAML+OIL are more recent efforts in the Semantic Web domain to combine the best features of RDF, SHOE, and OIL. They have well-defined semantics for representing axioms, conditions and constraints on the different entities that describe content. The discovery of relevant content in a Web page can also be achieved using Embedded Grammar Tags (EGT) (Gautham and Yacoob, 2002) instead of the usual tags. By this approach, once the information has been detected, a generative grammar that may correspond to an unforeseen user's query is also mapped. However, all these extensions still do not enable agents to extract only the desired response to a user's query. The general outline of the markup languages layer model for the Web is depicted in figure 1. The proposed framework, which for reasons of simplicity was named as *Bayesian XML* or *BXML,* lies above the layer of standard XML.

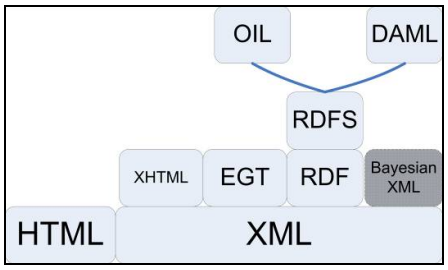

Figure1: The layer markup language model of the Web.

The principle idea is that the Web page contains text or parts of text that comprise the semantic meaning of a domain, or some part of it. For example, a page containing the following text: "*NASDAQ quote reached 2032 units*" may imply that the financial index was either dropped or raised compared to its yesterday quote. By using BXML, the author of a Web page can freely mark any words or phrases as useful input lexical items and can also annotate using a simple, yet effective and powerful specially designed annotation tool to provide the semantic categories that were triggered by the above lexical items. Note that these categories, in order to have a wide-ranging framework for other related Web pages have to be pre-ascertained by a domain expert.

## 2. Probabilistic Analysis of Semantic Interpretation

Semantic interpretation of an input query could be considered as the process of searching for the optimal (most probable) semantic interpretation through the space of candidate similar semantic interpretations for a specific domain, given the lexical items that define the meaning of the query. In the more general case, one would claim that the above mentioned hypothesis space actually contains all the semantic categories of the domain, however, during search process, those who do not resemble the candidate semantic interpretations are superseded.

In our approach, a stochastic model for modeling semantic disambiguation is defined over a search space $H*T$, where $H$ denotes the set of possible lexical contexts that could be identified within an input query $\{h_1,...,h_k\}$ or *input variables* and $T$ denotes the set of the allowable semantic interpretations of that question $\{t_1,\ldots,t_n\}$. Using Bayes' rule, the probability of the optimal interpretation $T_{opt}$ equals to:

$$T_{opt} = \arg\max_{T\in(t_1\ldots t_n)} p(T\mid H) = \arg\max_{T\in(t_1\ldots t_n)} \frac{p(H\mid T)p(T)}{p(H)} \rightarrow$$
$$T_{opt} = \arg\max_{T\in(t_1\ldots t_n)} p(H\mid T)p(T) \quad (1)$$

For a given observation sequence of input observations $\{h_1,...,h_k\}$, the above equation is modified into:

$$T_{opt} = \arg\max_{t_i\in(t_1,\ldots,t_n)} \frac{p(t_i)p(h_1,...,h_k\mid t_i)}{p(h_1,...,h_k)} \rightarrow$$
$$T_{opt} = \arg\max_{t_i\in(t_1,\ldots,t_n)} p(t_i)p(h_1,...,h_k\mid t_i) \quad (2)$$

The probability $p(h_1,...,h_k)$ is omitted since it remains the same for every $t_i \in \{t_1,...,t_n\}$, thus not affecting the *argmax* function. There are two possible assumptions that can be considered from this point, regarding how the lexical items are considered to be; either to be regarded as independent of each other or to take into account that there is some specific kind of dependency among all or a subset of them. If one assumes lexical independence, the naïve Bayesian classifier can be used. Nevertheless, the richness of the language often includes situations where certain words are used to denote a different meaning or to simply stress the sense of a particular word or phrase. Adjectives and adverbs are part-of-speech categories that generally modify the interpretation of a word or a phrase (usually the neighboring one). In such cases, Bayesian networks appear to be more suitable since they provide mechanisms for establishing a semantic-based representation of variables, a notion that is more human-centric than other, statistical methods.

## 3. Bayesian Networks

Bayesian networks provide a comprehensive means for effective representation of independence assumptions. They allow asserting conditional independence assumptions that apply to all or to subsets of the variables. A Bayesian network is consisted of a qualitative and quantitative portion, namely its structure and its conditional probability distributions respectively. Given a set of attributes $A=\{A_1,...,A_k\}$, where each variable $A_i$ could take values from a finite set, a Bayesian network describes the probability distribution over this set of variables. We use capital letters as $X,Y$ to denote variables and lower case as $x,y$, to denote values taken by these variables. Formally, a Bayesian network is an annotated directed acyclic graph (DAG) that encodes a joint probability distribution. We denote a network $B$ as a pair $B=<S,P>$ (Pearl, 1988) where $S$ is a DAG whose nodes correspond to the attributes of $A$. $P$ refers to the set of probability distributions that quantifies the network. $S$ embeds the following conditional independence assumption:

*Each variable $A_i$ is independent of its non-descendants given its parent nodes.*

$P$ includes information about the probability distribution of a value $a_i$ of variable $A_i$, given the values of its immediate predecessors in the graph, which are also called *parents*. This probability distribution is stored in a table, which is called conditional probability table. The unique joint probability distribution over $A$ that a network $B$ describes can be computed using:

$$p_B(A_1,...,A_n) = \prod_{i=1}^{n} p(A_i \mid parents(A_i)) \quad (3)$$

Taking into account equation (3), formula (2) can be re-written as:

$$T_{opt} = \arg\max_{t_j\in\{t_1,\ldots,t_n\}} p(t_j)\prod_{i=1}^{k} p(h_i \mid parents(h_i), t_j) \quad (4)$$

### 3.1 Learning Bayesian Networks from Data

In order to approximate the terms of equation (4), the structure of the network has to be provided. There are two practices for determining the structure of a Bayesian network. Either manually, by a human domain expert who should provide the interconnection of the variables, or having the structure determined automatically by learning from a set of training examples. Regarding the learning of the conditional probability table of a network, the same principle applies. The parameters of the table could either be provided manually by an expert or automatically

through a learning procedure. The task of manually supplying the parameters is a laborious one. Besides, in some applications it is simply infeasible for a human expert to know a priori both the structure and the conditional probability distributions. The problem of finding the most probable network structure from data is known to be NP-hard (Mitchell, 1997). The most commonly utilized approach is the introduction of a scoring metric that evaluates the probability of a candidate structure *B* over the training set *D*. The two standard metrics used to learn networks from data are the *Bayesian scoring function* (Cooper and Herskovits, 1992) and the one which is based on the principle of *minimal description length* (MDL) (Friedman, Geiger and Goldszmidt, 1997). Nevertheless, Heckerman (1995) observed that the two metrics are asymptotically equivalent as the sample size increases. Furthermore, they prove to be asymptotically correct, meaning that with probability one, the learned distribution converges to the underlying distribution as the number of training instances increases. For our approach, we used the former metric for determining the most probable network structure over a given training set.

## 4. Implementing the Bayesian XML Framework

Our approach focuses on empowering intelligent agents with NL understanding capabilities. Instead of using pre-defined, hand-coded grammars, we choose to establish a statistical framework, such as that of Bayesian networks, which depict probability distributions and concept relations in a graphical way, thus being more elaborate than other probabilistic representation machineries. Manual insertion of a priori grammar rules is cumbersome and cannot always efficiently cope with ill-formed sentences, such as those which contain misspellings or elliptical sentences. On the other hand, Bayesian networks can cope with such restrictions since the mapping from the lexical layer to the semantic layer is automatically performed using standard and evaluated machineries such as the Bayesian scoring function. The working model of the proposed framework is separated into two phases.

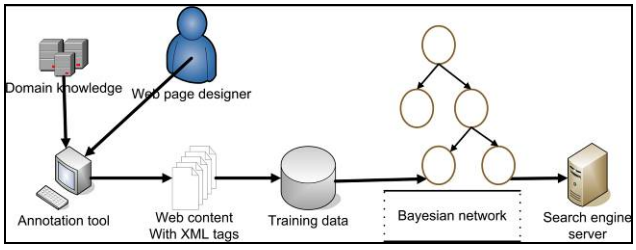


Figure2: The annotation stage of the proposed framework.

The former, depicted in figure 2, takes place at the Web page design time. More specifically, a human domain expert defines the semantic categories that can be reflected. It would be ideal to establish universal semantic categories and formalisms for common Internet areas in order for more pages to be integrated in an NL search. Upon completion of the domain knowledge definitions, the designer annotates those parts of the page that are considered to contain essential information. In order to make the process more easy, we have built an annotation tool (Maragoudakis, Fakotakis and Kokkinakis, 2004) that initially performs shallow parsing on a selected sentence and extracts the tuples of Subject-Verb-Object(direct and indirect) automatically. Subsequently, the annotator marks the keywords that characterize the meaning of the sentence and then maps this meaning by selecting the corresponding semantic category found in the tool. Recall that from a given set of possible domain semantic categories, only those who are actually affected are annotated. The tags are inserted to the page in an XML-like format, in order for Web crawlers to easily parse any pages that contain similar content and extract the instances used for Bayesian training. Such instances contain the lexical cases and the semantic tags. The trained network is encoded into the BXML search engine server for future NL searches.

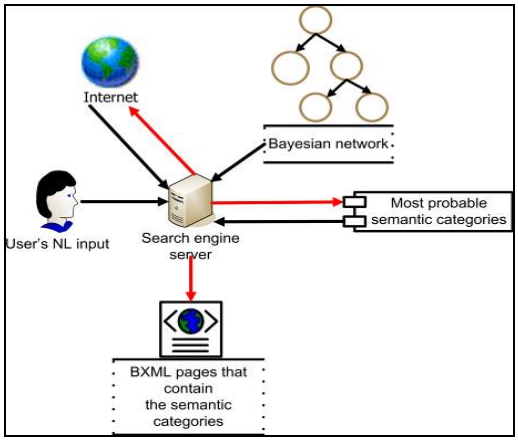


Figure 3: The NL information retrieval phase.

The latter phase corresponds to the run-time process. The NL query is quickly parsed in order to extract the significant parts of it such as nouns, numerical expressions, etc. The Bayesian network is consulted in order to infer on the most probable semantic categories, given the lexical items found in the user's query. Subsequently, the search engine looks for BXML Web pages that contain the predicted semantic categories and returns them to the user. Figure 3 outlines the above process.

### 4.1 Main Characteristics of Bayesian XML

Researchers have established a plethora of expressive Semantic Web languages that can be used to support accurate responses to NL queries. As an example, consider the RDF style in an imaginary 2004 Olympics' content:

Athens-*[hostof]* →*Olympic Games 2004*

Kenteris-*[worldrecordman]* →*200m*

Provided that agents embed these definitions, NL queries can be comprehended.

Using EGT markup language to Web content also allows for NL information retrieval. Consider again the same 2004 Olympics example:

*<info>*Kenteris*<EGT-in>who is the 200m world recordman</EGT-in>* lives in Athens *<EGT-in>Which city hosts the 2004 Olympics</EGT-in>* where he can train better.*</info>*

Despite the fact that annotation by using the above methods is quite straightforward, the resulted Web content is very specific, in the sense that the designer should fully

match an unpredicted query. In order to alleviate that problem, numerous different tags have to be added, covering the plentiful different kinds of NL queries that can be encountered. The update functionality is also circumscribed by the narrow grammar markup style. Imagine the human effort needed in order to label many pieces of a Web page that are considered as important with grammatical tags. The Bayesian XML search and markup offer three main advantages over the existing approaches.

- First, during annotation, the key lexical items are easily mapped to their concise semantic interpretation, bypassing the painstaking job of adding the entire anticipated NL queries.
- Second, the proposed framework significantly reduces the load of current search engines in the sense that the semantic interpretation of a Web page is included within. Furthermore, a portion of the representation is also included in the Bayesian networks of the domain.
- Third, we claim that update of Web content with new elements can be effortlessly achieved, provided that the domain semantic categories have already been established. Note also that there is a need for Bayesian XML Web crawlers to search for any changes of the Web pages, in order to periodically re-train the corresponding Bayesian networks.

## 5. Example Bayesian XML representation

As a first attempt to implement BXML, a medical domain was selected. In an already operational Web portal that provides users with information about a variety of medicines for the treatment of pneumonia, we applied the proposed formalism, in order to enhance the semantic meaning of the pages. The following illustration shows how a simple text entry can be augmented to embed semantic content using BXML.

The Cefaclor is harmless for pregnant women.

<*BXML*>The<***surface node***="***Active substance***"> Cefaclor<***/surface***>is<***surface node***="***Warnings***"> harmless<***/surface***> for <***surface node="Patient"***>pregnant women<***/surface***>

<***semantics Domain="Pneumonia"***><***/semantics***>

<***semantics Contraindications="No"***><***/semantics***>

<***semantics Period="Gestation"***><***/semantics***></*BXML*>

We annotated Web pages that contain information on about 50 pneumonia antibiotics, the parsing of which resulted in a total of 2500 training instances. The semantic category prediction performance of the Bayesian network on this set was estimated in a scale of 86%±2.3% using the 10-fold cross validation method. Furthermore, since we did not have other relevant Web content to deal with, we carried out a qualitative evaluation by introducing the system to 15 users which were supposed to provide 10 NL queries each. The objective for the system was to manage to find the most relevant semantic category that each query implied. Only one reformulation from the user's point of view was allowed. Table 1 exemplifies the outcome of these experiments.

| ***Category*** | ***Questions*** | **Error rate** |
|---|---|---|
| Initial queries set | 150 | |
| Reformulated queries | 32 | *21.3%(32/150)* |
| Unidentified queries after one reformulation | 12 | *37.5%(12/32)* |

Table 1: Query understanding performance of the proposed framework

As tabulated in table 1, from the initial set of 150 unanticipated queries the system achieved to estimate the correct semantic interpretation of 118 of them, resulting in a 21.3% error rate. 32 queries were reformulated and finally 12 of them were unable to be interpreted. Nevertheless, the assessment is still far from optimal since the domain was restricted to include a limited number of Web pages and only they constituted both the test and the training set. Before the BXML architecture is adopted as a Semantic Web markup representation, it is necessary to incorporate these essential characteristics: (a) annotation using more sophisticated automatic approaches such as linguistic tools; (b) improved adaptation of the structure and the parameters of the Bayesian networks in order to capture the training set better and (c) establishment of a common configuration for popular Internet domains, such as stock market, e-shopping, weather information, academic institutes, etc., so that more Web pages can incorporate the BXML semantic content.

## Bibliographical References


Soderland, S. (1997). Learning to Extract Text-Based Information from the World Wide Web, In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp. 251-254.

Fensel, D. (2000). The Semantic Web and Its Languages, IEEE Intelligent Systems, vol. 15, no. 6, pp. 67-73

Heflin, J., Hendler, J. & Luke, S. (1999). SHOE: A Knowledge Representation Language for Internet Applications, tech. report CS-TR-4078, Univ. of Maryland.

Gautham, D. & Yacoob, Y. (2002). Embedded Grammar Tags: Advancing Natural Language Interaction on the Web, IEEE Intelligent Systems, vol. 17, no. 1, pp. 48-53

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann.

Mitchell, T. (1997). Machine Learning, Mc Graw-Hill.

Cooper, J. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning, vol. 9, pp. 309-347.

Friedman, N., Geiger, D. & Goldszmidt, M. (1997). Bayesian network classifiers, Machine Learning, vol. 29, no. 2-3, pp. 131-163.

Heckerman, D., Geiger, D. & Chickering, D. (1995). Learning Bayesian Networks: the Combination of Knowledge and Statistical Data, Machine Learning, vol. 20, no. 3, pp. 197-243.

Maragoudakis, M., Fakotakis, N. & Kokkinakis, G. (2004). A Bayesian Model for Shallow Syntactic Parsing of Natural Language Texts, In Proceedings of the 4th International Conference on language resources and evaluation, Portugal, *to appear*.