Ontology-Based Binary-Categorization of Multiple-Record Web Documents Using a Probabilistic Retrieval Model

A Thesis Proposal

Presented to the

Department of Computer Science

Brigham Young University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Quan Wang

November, 2000

1 Introduction

The World Wide Web contains a tremendous amount of information. It is challenging to determine which Web documents are relevant to a user query, and even more challenging to rank them according to their degrees of relevance to the user query.

Much effort has been made to study different information retrieval (IR) approaches to solve these problems. Early IR models include the Boolean model and the Vector Space model (VSM) [BR99]. The Boolean model is based on binary decisions without a grading scale, and hence cannot be used for relevance ranking. VSM overcomes the limitation of the Boolean model by assigning non-binary weights to index terms in documents, as well as in user queries, and by computing the degree of similarity between a document and a user query using the assigned weights. Since there is an intrinsic limitation that these two models cannot handle any dependent relationships among index terms, they suffer from the problem of oversimplification. Retrieved evidence (e.g. keywords or index terms) is assumed to be mutually independent and neither structure (i.e. chapters, sections, etc.) nor semantics (e.g. closely-related index terms) that appear in documents to be examined or in user queries to be processed are considered.

The probabilistic model, on the other hand, solves these IR problems by applying probability theory. In this thesis, we consider applying probability theory for categorizing a collection of documents into two non-overlapping sets according to an application ontology[1]. Documents in one set are relevant to the application ontology, whereas documents in the other are not.

In order to automate the binary-categorization process, we must make a good guess based on imperfect knowledge about the documents. We consider an approach that establishes a relevance probability of a document with respect to an application ontology by *curve-fitting* over relevance probabilities. Figure 1 shows a sample curve for relevance probability fitting with respect to the term frequencies of an index term. Since the probability curve looks like an "S", it is called an *S-curve*. Our approach studies the relationship between a relevance

[1] An application ontology is a conceptual model instance that describes a real-world application of interest [EJN99].

probability and a set of <index term, term frequency (tf)> pairs and tries to estimate a relevance probability based on the probability-tf curves. In addition, we adopt two out of three heuristics proposed in [ENX00], namely, (1) a *density heuristic* that measures the percentage of index terms in a document which match the lexical values specified in an application ontology, (2) a *grouping heuristic* that measures the degree to which the potential attribute-value pairs can be grouped as ontology records.

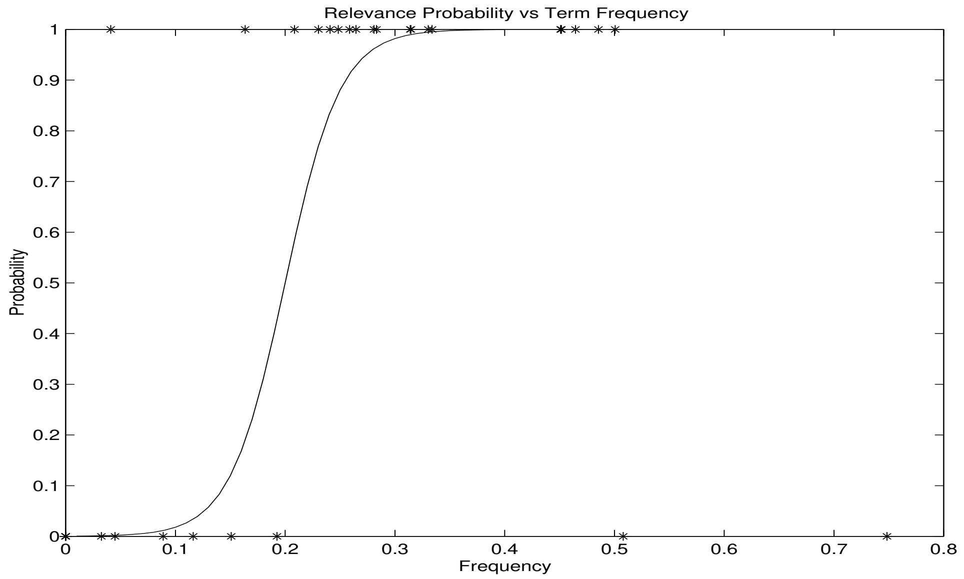

Figure 1: A relevance probability fitting curve

In our categorization approach, we adopt *regression analysis*, a highly successful statistical tool, to study the relationship between relevance probabilities and term frequencies of the index terms, the density and the grouping value and construct (i.e. fit) the relevance probability curve using documents in a training set. Since the relevance probabilities of a given set of documents are discrete[2], a logistic function[3] should be the most ideal choice for fitting the probability curve, and hence we choose logistic regression over other regression analyses. Once we have a probability curve such as the one shown in Figure 1, we can look at

[2]Documents in a given set are divided into two groups. The documents in one group are relevant to the ontology and each one in the group has the relevance probability of 1, whereas documents in the other group are not relevant and are assigned the probability of 0.

[3]A logistic function is a *continuous function* bounded between 0 and 1 with $y = 0$ and 1 as the asymptotes.

the curve to estimate the corresponding relevance probability of the document with respect to an application ontology.

A benefit of using logistic regression analysis to estimate a relevance probability is that we can analyze any dependent relationships among index terms. As mentioned earlier, some IR models can not handle any dependent relationships among index terms, and hence suffer from oversimplification. The *binary independence indexing model* and the *binary independence retrieval model*, for example, use mutual independence assumptions exclusively to make the models computationally workable, which by no means provides any justification to do so and inevitably causes distortion. Unlike many other probabilistic retrieval models that have independence assumptions and fail to exclude or compensate for the resultant systematic biases, our probabilistic model tries to weaken the independence assumptions and is capable of analyzing and compensating for existing dependence relations by using logistic regression analysis.

In this thesis, we will adopt a probabilistic retrieval model using logistic regression for recognizing multiple-record Web documents against an application ontology. A document is categoried as a *multiple-record* document if it contains a sequence of similar chunks of information about the main entity of an ontology [ECJ99]. The proposed probabilistic model will recognize multiple-record Web documents that are relevant to a chosen application of interest defined as an application ontology. We have performed some preliminary experiments using the proposed probabilistic retrieval model. The results are promising. The proposed probabilistic approach based on logistic regression analysis does better, or at least as good as other IR models, such as the Vector Space model [Kwong99] and the heuristics approach [ENX00] on the same training set and test set compiled for [ENX00].

2 Thesis Statement

The purpose of this thesis is to study the applicability of our proposed probabilistic information retrieval model based on logistic regression analysis for recognizing multiple-record Web documents with respect to a particular application ontology. Our approach

1. analyzes data extracted from a training set of multiple-record Web documents with respect to an application ontology and sets up a logistic regression analysis model that determines the *probability fitting curves* using a set of index terms extracted from the documents in the training set;

2. calculates the *relevance probability* of a given document with respect to an application ontology based on the probability curve-fittings obtained in (1);

3. studies dependent relationships among index terms and includes important ones in our *relevance probability estimation*; and

4. evaluates the performance of our probabilistic retrieval model in terms of *precision, recall*, and *accuracy* measurements.

3 Methods

In this thesis we will estimate the relevance (R) of a document d with respect to an application ontology q by calculating the relevance probability $p_q(R|d)$ using logistic regression analysis. (Whenever it is clear which application ontology we use, we can simply rewrite the relevance probability symbol as $p(R|d)$.) To estimate the relevance probability of d with respect to q, we make the following assumptions:

1. The relevance of d to q depends only on d and q. It is independent of other documents in the collection to which d belongs.

2. Based on an application ontology, each document in the collection is described by:

   (a) a set of <index terms : term frequency> pairs $A_1 : x_1, \cdots, A_n : x_n$, $n \geq 1$.

   (b) a *density heuristic* value, which measures the percentage of a document applicable to the application ontology.

   (c) a *grouping heuristic* value, which measures the degree of structural similarity between chunks of information in a document and application-ontology records.

The first assumption simply states the fact that each document retrieval is an independent event; the relevance of documents determined earlier will not affect the relevance of others to be determined later. Part (a) in the second assumption states that we abstractly view a document d as $(x_1, \cdots, x_n)$ where the term order is assumed to be fixed as $A_1, \cdots, A_n$, such that $p(R|d)$ can be written as $p(R|(x_1, \cdots, x_n))$. Parts (b) and (c) will be used to improve the accuracy of our probabilistic retrieval model. Our proposed approach estimates $p(R|(x_1, \cdots, x_n))$ in two stages. In the first stage, $p(R|(x_1, \cdots, x_n))$ is broken down into $p(R|x_1), \cdots, p(R|x_n)$ using an independence assumption. In the second stage, dependent relationships among index terms are studied and included in the relevance probability estimation to compensate for any bias caused by the assumptions made in the first stage.

Relevance probabilities over the term frequencies of an index term for a set of documents is usually an S-curve, and thus each $p(R|x_i)$ is best represented by the following logistic curve fitting function.

$$p(R|x_i) = \frac{1}{1 + \exp(-(C_0 + C_1 x_i))} \tag{1}$$

In Equation 1, x_i is the term frequency of index term A_i. Since one of the most commonly-used measure of an index term is its term frequency, we only use the term frequency of an index term in Equation 1. The relevance probability $p(R|(x_1, \cdots, x_n))$ is then calculated by multiplying all the $p(R|x_i)$, $1 \leq i \leq n$, where n is the total number of index terms used in the determination of relevance probabilities.

In Equation 1, the unknown coefficients C_0 and C_1 are called *regression coefficients*. They determine the exact form of Equation 1. In order to calculate C_0 and C_1, we will construct a set of training documents whose relevance probabilities with respect to a particular application ontology are known to be either 1 or 0. Since the term frequencies for all the index terms A_i $(i = 1, \cdots, n)$ can be extracted from the training set, we construct for each chosen index term A_i a list of <relevance probability, term frequency> pairs (i.e. $< p(R|x_i),\ x_i >$) from all the documents in the training set. The logistic regression software package takes this list of ordered pairs as an input and generates a pair of regression coefficients C_0 and C_1 as output.

After we have determined the regression coefficients C_0 and C_1, we can use the same curve-fitting function given in Equation 1 to calculate the probability $p(R|x_i)$, $1 \leq i \leq n$, for each document in a test set. First, the term frequency x_i of the index term A_i is extracted from the document, and its value is substituted into Equation 1 to evaluate $p(R|x_i)$. The same procedure is repeated for all the other index terms A_j $(1 \leq i, j \leq n,\ i \neq j)$. The relevance probability $p(R|(x_1, \cdots, x_n))$ is then estimated by multiplying all the $p(R|x_i)$, $1 \leq i \leq n$. The density and grouping values are treated the same as term frequencies of index terms such that $p(R|Density)$ and $p(R|Grouping)$ are estimated using the same logistic regression procedure.

When we perform logistic regression analysis for each chosen index term A_i $(1 \leq i \leq n)$, we not only compute the regression coefficients C_0 and C_1, but also obtain the statistical information, the *p-value*[4], associated with it. The *p-value* indicates (i) how significant the corresponding index term is in the determination of the relevance probability of a document

[4] A p-value is a risk factor, which is usually assigned a value of 0.05 (i.e., 95% confidence) [HL89] to assess the significance of a chosen index term. If the *p-value* of an index term is smaller than the risk factor, it is considered to be significance.

with respect to the corresponding application ontology and (ii) how confident we are in these statistical results. This information will serve as a guide for choosing only significant index terms when we construct the proposed probabilistic retrieval model.

In the next step, we will use the *p-value*s of index terms to determine which corresponding significant index terms should be included in computing the relevance probabilities. Adding statistically insignificant index terms can potentially cause over-fitting and cause the performance of our probabilistic model to deteriorate. Oversimplification, such as the absolute independence assumption and the conditional independence assumption, however, can give rise to computational efficiency but it comes with an expense in accuracy. An ideal situation is to discard any independence assumption, even though it is almost mathematically intractable to do so. Therefore, we have to consider the tradeoff between computational efficiency and accuracy. In practice we can expect to achieve more accuracy of relevance probability estimation without significantly increasing the computational complexity. This can be accomplished by first using as few independence assumptions as possible and then by using a *correlation function*[5] to identify any significant dependent relationships among index terms and including the significant ones in the probability calculation.

In order to have a robust probabilistic retrieval model, it is necessary to collect the training and test documents that are both narrow and broad in content. The content should be *broad* because we want our probabilistic model to handle different kinds of documents. The content should be *narrow* because we would like to catch subtle differences.

The performance of our probabilistic model will be evaluated by calculating *precision*, *recall*, and *accuracy* based on a set of test documents. Precision is the proportion of retrieved documents actually relevant, recall is the proportion of relevant documents actually retrieved, and accuracy is a measurement which combines the precision and recall and evaluates the performance of the proposed model by giving the ratio of the total number of correct document recognized to the total number of documents processed.

[5] A correlation function is used to study the dependent relationship between two quantities and is usually defined as the product of the two quantities.

4 Thesis Schedule

A tentative schedule of this thesis is as following:

Literature search and reading	May 01, 2000
Chapter 2, 3, 4	October 31, 2000
Chapter 5	November 30, 2000
Chapter 1, 6	December 6, 2000
Thesis review and defense	December 20, 2000

5 Thesis Outline

1 Introduction

1.1 The problem

1.2 Thesis organization

Estimated length : 3 pages

2 Application Ontology

Estimated length : 2 pages

3 Logistic regression Analysis

3.1 Regression analysis

3.2 Logistic regression analysis

Estimated length : 10 pages

4 Probabilistic retrieval Based on Logistic Regression

4.1 Relevance assumptions

4.2 Algorithms to calculate the relevance probability

Estimated length : 10 pages

5 Analysis of the Probabilistic Retrieval Model

5.1 Experimental results

5.2 Study the impact of the independence assumptions

Estimated length : 10 pages

6 Conclusion and future work

Estimated length : 2 pages

6 Delimitations of the Thesis

This thesis will not do the following:

(1) deal with Web documents other than multiple-record Web documents

(2) prove the adopted technique is superior to the other methods proposed

7 Contribution to Computer Science

The contribution of this thesis includes

(1) providing a probabilistic model for information retrieval which is applicable for multiple-record Web documents

(2) studying the dependent relationships among index terms such that we can accomplish a balanced tradeoff between computation complexity and retrieval accuracy.

8 Bibliography

[AA84] A. Albert, J.A. Anderson. On the Existence of Maximum Likelihood Estimates in Logistic Regression Model. *Biometriks*, 71, (1), Page 1-10, 1984.
This paper discusses the problem of the existence of maximum likelihood function in logistic regression analysis. It categorizes the data patterns into three mutually exclusive groups, complete separation, quasicomplete separation and overlap. It concludes that the difficulties associated with complete separation and quasicomplete separation are caused by small sample size. We have encounted the divergent problem mentioned in this paper when we did logistic regression analysis. The paper proposes general ways for identifying nonconvergence and suggests methods to avoid infinite coefficient problem.

[BJ81] Maurice C. Bryson and Mark E. Johnson. The Incidence of Monotone Likelihoods in the Cox Model. *Technometrics*, Vol. 23, No. 4, November 1981.
This paper addresses the problem of infinite maximum likelihood estimate, such as unconvergent regression coefficients which are likely to happen during logistic regression analysis. It proposes ways to detect the monotone likelihood problem and suggests steps to avoid and attack the problem if nonconvergence exists.

[BR99] Richardo Baeza-Yates, Berthier Riberro-Neto. *Modern Information Retrieval.* Addison-Wesley, 1999.
This book covers the details of information retrieval.

[CGD92] Williams S. Cooper, Fredric C. Gey and Danial P. Dabney. Probabilistic Retrieval Based on Staged Logistic Regression. *15th Ann Int'l SIGIR*, 198-210, 1992.
This paper suggests a scheme for probabilistic retrieval based on staged logistic regression analysis. The method proposed in this paper tries to weaken the independence assumptions made by most probabilistic retrieval models and compensate for the bias effect of the assumptions by adding extra regression stages. This thesis will adopt the basic approaches outlined in this paper.

[CLR98] Fabio Crestani, Mounia Lalmas, Cornelis J. van Rijsbergen, and Iain Campbell. "Is this document Relevant? $\cdots$ Probably". A survey of Probabilistic Models in Information Retrieval. *ACM Computing Surveys*, Vol. 30, No. 4, 528-552, December 1998.
This paper reviews the history of the probabilistic models in information retrieval. It describes the strengths and the problems of the probabilistic retrieval methods. The author gives a survey of several popular probabilistic retrieval models.

[CS89] D.R. Cox and E.J. Snell. *Analysis of Binary Data.* Chapman and Hall, 1989.
This book provides the theoretical background of logistic regression in analysis of binary data. It emphasizes in providing theoretical justification of logistic regression in dealing with binary data.

[ECJ99] D. W. Embley, D. M. Campbell, Y. S. Jiang, W. Liddle, D. W. Lonsdale, Y.-K. Ng and R. D. Smith. Conceptual-model-based Date Extraction from Multiple-Record Web Pages. *Data and Knowledge Engineering*, 31(3), 227-251, Nov. 1999.
This paper proposes a conceptual-modeling approach to extract and structure data automatically. The data of multiple-record Web documents needed in this thesis will be extracted from the software package constructed using the schemes outlined in this paper.

[ECSL98] David W. Embley, Douglas M. Campbell, Randy D. Smith, Stephen W. Liddle. Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. *Proceedings of CIKM*, 52-59, 1998.
This paper proposes a new approach to extract information from unstructured documents based on application ontologies. The approach targets the documents which are data-rich and narrow in ontological breadth, which will be the only type of documents considered in this thesis.

[EFKR99] David W. Embley, Norbert Fuhr, Claus-Peter Klas, Thomas Rolleke. Ontology Suitability for Uncertain Extraction of Information from Multiple-Record Web Documents.

Databank Runbrief 24:48-53, 1999.
This paper introduces the idea of using probabilistic approach to determine ontology suitability of multiple-record Web documents. It proposes the idea which embeds three heuristics, namely density, expected-values and grouping in probability frame. The paper does not give a detailed description about how to implement the model, instead it outlines some potential possibilities but still leaves rooms for other approaches such as logistic regression analysis adopted in this thesis.

[EJN99] D.W. Embley, Y.S. Jiang, and Y.-K. Ng. Record-boundary discovery in Web document. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD'99)*, page 467-478, Philadelphia, Pennsylvania, June 1999.
This paper proposes a heuristic approach to chunk unstructured or semistructured Web documents that contain multiple records according to record boundaries. In this approach, the structure of a Web document is viewed as a tree of nested HTML tags and the records of interest are located using five independent heuristics.

[ENX00] D.W. Embley, Y.-K. Ng, L. Xu. Filtering Multiple-Record Web Documents Based on Application Ontologies. (submitted for publication)
This paper proposes an approach to filter multiple-record Web documents based on application ontologies. The filtering scheme uses machine-learning algorithm over there heuristics (density, expected-values, grouping) to determine the relevance of multiple-record Web documents to the ontology. Among three heuristics, the expected values are evaluated using the Vector Space model

[Fuh92] Norbert Fuhr. Probabilistic Models in Information Retrieval. *The Computer Journal* 35(3):243-255, 1992.
This paper gives an introduction and survey of probabilistic information retrieval models. It also discusses some common problems of probabilistic retrieval models.

[GEY94] Fredric C. Gey. Inferring Probability Relevance Using the Method of Logistic Re-

gression. *Proceedings of SIGIR*, 222-231, 1994.
This paper investigates a probabilistic inference system based on logistic regression analysis. The proposed method was tested on three well-known information retrieval test collections. The experimental results show that the proposed method performs much better or as well as the other retrieval models, such as the Vector Space model.

[HL89] David W. Hosmer, Jr. and Stanley Lemesshow. *Applied Logistic Regression*. John Wiley and sons, 1989.
This book introduces the logistic regression model. It covers the introduction to the logistic regression model and the model-building strategies and methods for logistic regression.

[Kwong99] L.W. Kwong. Multiple-Record Web Documents Based on Application Ontologies and Information Retrieval. Master thesis, Computer Science Department, Brigham Young University, July 1999.
This paper proposes an automated way to filter multiple-record Web documents based on application ontologies and the Vector Space model (VSM). The paper also extends VSM to include the magnitudes together with the directions in decision making rule.

[HPS96] David A.Hull, Jan O.Pedersen, Hinrich Schütze. Method Combination For Document Filtering. *Proceedings of SIGIR*, 279-287, 1996.
This paper introduces an idea to improve performance of document filtering by combining several retrieval methods. The paper suggests that it is a good strategy to develop statistical classifiers that produce accurate estimations of relevance probability, as well as to seek method to combine these estimates. The paper proposes several combination strategies and provides some strategy selection rules and the justification, such as combinations using logistic regression and machine learning algorithm.

[RIJ75] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, 1975.
This book contains basic concepts of information retrieval. Its Chapter 6 outlines fundamental principles used in many probabilistic information retrieval methods.

[TPL95] Markus Tresch, neal Palmer, Allen Luniewski. Type Classification of Semi-Structured Documents. Proceeding of the 21th VLDB Conference.
WWW contains a huge amount of unstructured or semistructured documents such as journal articles, Email and HTMLs. These document are often not explicitly typed, however they must be assigned a explicit type in order to apply object-oriented technology. This paper presents a vector space classifier for determining the type of semi-structured documents. It also introduces three basic criteria for any document classification, which are high accuracy, fast classification and incremental extensibility.

[Trivedi82] Kishor Sheridharbhai Triveedi. *Probability and Statistics with Reliability, Queuing and Computer Science Applications.* Prentice-Hall, Inc., 1982.
This book introduces the fundamental concepts and theories of probability and statistics.

9 Artifacts

The logistic regression analysis will be performed using the statistical software package **SPSS**. The program which implements our probabilistic retrieval model will be written in C^{++} and **Matlab** on Window-based PC systems.

10 Signatures

This proposal, by Quan Wang, is accepted in its present form by the Department of Computer Science of Brigham Young University as satisfying the proposal requirement for the degree of Master of Science.

Yiu-Kai (Dennis) Ng, Committee Chairman

David W.Embley, Committee Member

Kent E.Seamons, Committee Member

J. Kelly Flanagan, Graduate Coordinator