A Novel Transcription Cataloging technique based on Combining Approach of Established Algorithms and Methodology for Deriving Unambiguous Results

Harish V. Gorewar
IV Semester, MTECH, CSE,
Sagar Institute of Research and Technology,
Bhopal (India)

Dr. M. Kumar
Department of Computer Science and Engineering
Sagar Institute of Research and Technology,
Bhopal (India)

ABSTRACT

This paper discusses a new approach for Transcription Cataloging (TC) based on combining efficient algorithms. The important aspect of automatically sorting and classifying a set of documents into any category by incorporating a predefined set is Transcription Cataloging. Automated Transcription Cataloging is gaining notability since it frees organizations from the hectic and time consuming need of manually organizing documents, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. In terms of accuracy, modern Transcription Cataloging systems proves better than that of trained human professionals, which is made possible by a combination of information retrieval technology and machine learning technology in Transcription Cataloging approach. There are numerable useful applications of this approach spanning various scientific and general fields of work. This paper deals in depth the feasibility of Transcription Cataloging pertaining to various domains along with making substantial use of techniques like document indexing, text filtering and classifier learning technique. Also the approaches of standard input and tokenization are considered for a better output which shall be devoid of any complexity for Transcription Cataloging.

General Terms

Information Retrieval, Knowledge Extraction

Keywords

Transcription cataloging, Fast-KNN & Naïve Bayesian algorithms, pair-relation, deductive inference.

1. INTRODUCTION

There is a growing body of research addressing automatic transcription cataloging. Probabilistic model in the work of Lewis [1] uses Bayesian independent classifiers for categorization. To cater to the need of classifying documents in a specific category, some form of cataloging of this textual information is required. Masand et al. [2] adopt a memory-based reasoning strategy to classify news stories.

After k best documents are retrieved, the weight of the associated categories is obtained by summing similarity scores from the near matches. Nevertheless, the magnitude of the number of documents of potential interest to a human classifier far exceeds to the magnitude of documents required to be classified. In this paper, we bring to forefront a way that can be very easily and commonly be used to limit the manual efforts in classifying documents to relevant topics by judiciously using software that does this work properly [2,6].

In view of the fact that, English language is very vast and global in nature and kind of sentences that can be made to convey one single fact are highly improbable, this research focuses on the reduction in ambiguity in categorization. This demands an immensely searched dimensional space so as to correctly associate a word to its appropriate category [7]. Although many approaches are proposed to cater to the needs of text categorization, they do not reach to the optimum level of accuracy and do possess few of the ambiguities in them. An important endeavor made in this research is to bring down the level of ambiguity in Transcription Cataloging considerably and enhance likelihood of accurate document categorization.

World contains enormous volume of unstructured data that is difficult to manage and utilize in a worthy way. It is impossible to provide services based on those unstructured data without certain level of cataloging and processing [3]. Transcription Cataloging is one of the most optimistic solutions to address this issue and has become an active research topic in Information Retrieval and Knowledge Extraction.

The Transcription Cataloging methods being discussed are completely general, and do not depend on the availability of special-purpose resources that might be unavailable or costly to develop [6]. These assumptions need not be verified in operational settings, where it is legitimate to use any source of information that might be available or deemed worth developing. Relying only on endogenous knowledge means classifying a document based solely on its semantics, and given that the semantics of a document is a subjective notion. This paper thus embarks on to take a closer look at a simplified Transcription Cataloging approach by describing the standard methodology through which a TC system is built [8].

2. BASICS OF TRANSCRIPTION CATALOGING

TC may be formalized as the task of approximating the unknown target function Φ : D × C → {T,F} (that describes how documents ought to be classified, according to a supposedly authoritative expert) by means of a function ˆΦ : D×C → {T,F} called the classifier, where C = {c1, . . . , c|C|} is a predefined set of categories and D is a (possibly infinite) set of documents. If Φ(dj, ci) = T, then dj is called a positive example (or a member) of ci, while if Φ(dj, ci) = F it is called a negative example of ci. The categories are just symbolic labels: no additional knowledge (of a procedural or declarative nature) of their meaning is usually available, and it is often the case that no metadata (such as e.g. publication date, document type, and publication source) is available either. In these cases, classification must be accomplished only on the basis of knowledge extracted from the documents themselves [4, 5]

Assumptions we follow are:

The categories are just symbolic labels, and no additional knowledge (of a procedural or declarative nature) of their meaning is available.

No external or exogenous knowledge (i.e., data provided for classification purposes by an external source) is available. In such cases, classification must be carried out on the basis of internal or endogenous knowledge only (i.e., knowledge extracted from the documents).In particular, this means that metadata such as, for example, publication date, document type, publication source, etc., is not assumed to be available.

- Document Organization

Performing the task of indexing using a controlled vocabulary can be seen as an instance of the general problem of document base organization. In general, many other issues pertaining to document organization and filing, be it for purposes of personal organization or structuring of a corporate document base, may be addressed by Transcription Cataloging techniques.

- Text Filtering

Text filtering is the activity of classifying a stream of incoming documents dispatched in an asynchronous way by an information producer to an information consumer. A typical case is a news feed, where the producer is a news agency and the consumer is a newspaper. In this case, the filtering system should block the delivery of the documents the consumer is likely not interested in (e.g., all news not concerning sports, in the case of a sports newspaper).

3. RESEARCH METHODOLOGY EMPLOYED

Effective computer generated classification solutions obviously increases efficiency and productivity. A computer can effectively process information in much faster way than humans. With the enormous growth of electronically stored text, efficiency is of profound importance. Beyond the immediate efficiency gains, however, is the great promise of machines that appear to be ‘reading’, machines that examine free text and make correct decisions. This paper put forward techniques that are currently feasible, that they are capable of processing huge numbers of documents in reasonable times, and that high performance is achievable when high quality sample data are available. We now discuss some of the actual techniques for dealing with the problems of document indexing and classifier learning.

Utilization of Naïve Bayesian algorithm was successfully carried out by Joachims [11] where it was probabilistically devised. Bayesian Network concept was made use of in effective manner by Sahami [10] in hierarchical document classifications. Weiss [10] incorporated Decision Tree concept in formulating rules that enhances in decision making in categorization. Human learning issues and its adaptation in Transcription Cataloging using Neural Network is devised through research by Yang [4] and it exhibits rational results along with Linear Regression and KNN [11]. A comprehensive comparative evaluation of a wide-range of Transcription Cataloging methods is reported in [3], [12] and [13].

Sun et al. [9] reviewed several concepts which serve as reference material for this work. Efficient implementation of existing algorithms for Transcription Cataloging and their usage individually or in combination and approximations of kernels for more efficient computation [14]. The availability of efficient cataloging methods and approximation schemes makes development of novel methodologies, especially suited for text comparison, an area especially stimulating and amenable to yield results which can prove useful in other domains, such as bioinformatics and multimedia retrieval.

4. RELATED WORK IN TRANSCRIPTION CATALOGING

TC was earlier conducted in innovative work on probabilistic text classification [12]. Since then, Transcription Cataloging approach has been used for a number of different applications. Note that the borders between the different classes of applications listed here are fuzzy and somehow artificial, and some of these may be considered special cases of others. Other applications we do not explicitly discuss are speech cataloging by means of a combination of speech recognition and TC.

4.1. Information Retrieval Techniques

TC heavily relies on the basic machinery of Information Retrieval. The reason is that TC is a content -based document amazement task, and as such it shares many characteristics with other IR tasks such as text search. IR techniques are used in three phases of the text classifier life cycle:

- IR-style *indexing* is always performed on the documents of the initial corpus and on those to be classified during the operational phase.
- IR-style techniques (such as document-request matching, query reformulation, are often used in the *inductive construction* of the classifiers.
- IR-style *evaluation* of the effectiveness of the classifiers is performed.

4.2. Rule-Based Cataloging Models

Machine learning systems solve problems by examining samples described in terms of measurements or features. For the application of machine learning methods, the samples of documents must be transformed into this type of representation. For transcription cataloging, an adaptation of a machine learning method must consider the following main processes:

- A preprocessing step for determining the values of the features or attributes that will used for representing the individual documents within a collection. This is essentially the dictionary creation process.

- A representation step for mapping each individual document into a training sample using the above dictionary, and associating it with a label that identities its category.

- An induction step for finding patterns that distinguish categories from one another.

The first step is to produce a list of attributes from samples of text of labeled documents, the dictionary. The attributes are single words or word phrases. Given an attribute list, sample cases can be described in terms of the words or phrases found in the documents. Each case consists of the values of the attributes for a single article, where the values could be either Boolean, e.g., indicating whether the attribute appears in the text or does not, or numerical, e.g., frequency of occurrence in the text being processed.

4.3. Text Representation

Document retrieval systems are supposed to choose documents that are about some concept of interest to the retriever. However, documents do not have concepts, but rather words. Words clearly do not correspond directly to concepts. Some words are used for more than one concept, e.g., \bank" as a financial institution and \bank" as part of a river. Some concepts require more than one word for their designation, e.g., the football player \running back," and most concepts can be referenced by more than one word or phrase, e.g. \doctor" and \physician." Humans are relatively good at inferring concepts from the words of a document. To do this, they bring to bear vast knowledge of the grammar of the language and of the world at large. Very little of this knowledge is available to a computer system, in large part because we have only sketchy and incomplete methods for organizing or inferring such information automatically.

4.4. Proposed Methodology

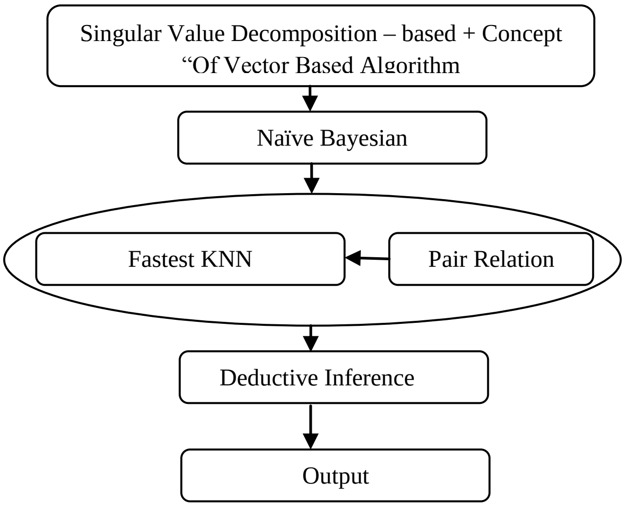
Singular Value Decomposition – based + Concept "Of Vector Based Algorithm
Naïve Bayesian
Fastest KNN
Pair Relation
Deductive Inference
Output

Figure 1: Working Flow of the proposed Transcription Cataloging system

This research adapts and incorporates different algorithms, few of which are notably used across the globe for the purpose of Transcription Cataloging and some are being developed through a comprehensive research and analysis. The approaches of Singular Value Decomposition, Vector Based algorithm [3], Naïve Bayesian probability algorithm [15], Fast - KNN technique [7], a novel approach of Pair Relations and Deductive Inference mechanism are incorporated and effectively used in the proposed system. A balanced use of a novel method of pair relation is introduced which brings in lot of innovative changes in the text to be categorized. Pairs are formed in accordance to their relation in the existing categories. If the relation is established, then a weight assignment process takes place which assures that the formed pairs does not get into any ambiguity further. Moreover, a process of deductive inference is incorporated which assumes all those important work which shall convey the meaning of the overall textual content. The detail insight is provided in this transcript and each of them is dealt with elaborations.

5. NAÏVE BAYES CATEGORIZATION ALGORITHM

Naïve Bayesian (NB) algorithm is one of the most widely used algorithms for document classification and it has been producing considerable results [15]. Naïve Bayesian algorithm computes the Posterior probabilities that the document belongs to different classes and assigns it the class with the highest posterior probability. The posterior probability of the class is computed using Bayes rule and the testing sample is assigned to the class with highest posterior probability. The novel part of the NB algorithm is the assumption of word's independence that the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given in that category. The fundamental notion is to use the joint probabilities of words and categories to estimate the probabilities of categories given in a document.

This paper makes use of one of the two versions of NB algorithm. That is, Multi-Variate Bernoulli Event Model [4]. This model takes into account only the presence or absence of a particular term in the given input. It is no way concerned with the number of occurrences of each word. It is worth noting that this thesis incorporates this idea to get to understand whether the word being received as an input is present in any of the categories or not. Furthermore, this algorithms' mechanism formulates in terms of probability features but deviating from probabilistic model, this thesis work uses in other way. It considers values that are pre assigned to each word and this value is in turn acts like a feature to establish posterior probability. This facet helps in determining the proximity on a word with a category.

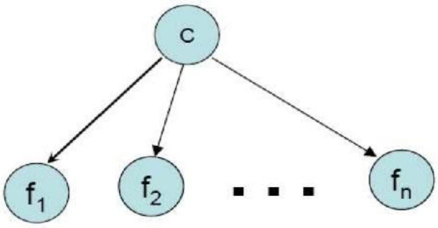

Figure 2: Probability feature of NB algorithm

Assumption: each fi is conditionally independent from fj given C.

Choose c* = arg maxcP(c) Π kP(fk| c) Two types of model parameters:

- Class prior: P(c)
- Conditional probability: P(fk| c)
- The number of model parameters:

|C|+|CV|

The algorithm is also widely used in text categorization. The mathematic method describes as follows: Calculate the probability vector (w1, w2, w3,….wn) of the characteristic word that belongs to every class:

Wk= P(Wk/Cj) = 1 + ∑i=1 N (Wk, di) / ∣V∣+ ∑s=1 V ∑i=1D N (Ws, di)

Then, Classify words according to the characteristic word as new text comes, then calculate the probability of text i d belonging to class Jc. Then, Compare the probability of new text belongs to each class and distributes the text to the class with the max probability.

6. SYSTEM WORKFLOW

The KNN algorithm applied to transcription cataloging is a simple, valid and non-parameter method. The traditional KNN has a fatal defect that the time of similarity computing is huge. The practicality will be lost when the KNN algorithm is applied to transcription cataloging with the high dimension and huge samples. In this paper, a method called TFKNN(Tree- Fast-K-Nearest-Neighbor) [7] is used along with other established algorithms and techniques, which can search the exact k nearest neighbors quickly. The KNN is a high performance classifier for text categorization. However, it is sensitive to high dimensional data. KNN classifier is an instance based learning algorithm that is based on a distance function for pairs of observations such a cosine distance. In this classification pattern, k nearest neighbors of a given data is computed first. Then the similarities of one particular sample from the testing data are aggregated to the k nearest neighbors according to the class of the neighbors. And then the sample being tested is assigned to the most similar class.

7. CONCLUSION

In this paper it is very aptly demonstrated how the innovative amendments and modifications in algorithms of Fast-KNN, Naïve Bayesian, Singular Value Deduction and Vector Based with addition of Pair Relation Method and Deductive Inference Technique can bring very significant and trend setting changes in the field of Text Categorization. These noticeable improvements steer the proposed approach towards an effective and comprehensive Transcription Cataloging system that can match user's perceived interests. The novelty of this method is that it can be worked out with lot of ease. As rich and comprehensive domain Transcription Cataloging systems are in place with complexities, the approach proposed in this thesis may be a suitable alternative to the traditional Transcription Cataloging methods. The strength comes from the substantial use of procedures and elaborative structures. Its precision measure is very reasonable and fetch out good results.

8. REFERENCES

[1] D.D. Lewis, "Featurc Sclection and Featurc Extraction fur Tcxt Categorization," Proc. Speech nnd Nniurnl LniipngP Workshop, pp.212-217, Arden House, 1992.

[2] B. Mesand, G. Linoff, and U. Waltz, "Classifying News Stories Using Memory Based Reasoning," Proc. 15th Int'l ACM SIGlR Coil/. Resenrcii and Development iii lnJorinntion I<etrievd, pp. 59-65,1992

[3] Bingheng Yan, Depei Qian (2007), "Building a Simple and Effective Text Categorization System using Relative Importance in Category", Third International Conference on Natural Computation (ICNC 2007), IEEE, vol-9, pp. 952-978.

[4] Yang Y (1994), "Expert Network: Effective and efficient learning from human decisions in text categorization and retrieval," Proc. of the seventeenth Int'l ACM SIGIR Conf. on Research and Devp. in information Retrieval, Dublin, pp.13-22.

[5] Salton. G., Wong. A. and Yang (1975),"A Vector Space Model for Automatic Indexing", Communications of ACM, vol 18, No. 11, pp. 613–620.

[6] Cohen W, Singer Y. (1996), "Context-sensitive learning methods for text categorization," Proc. of the 19th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval, Zurich, pp. 307~315.

[7] Yu Wang; Zheng-Ou Wang;(Aug. 2007) , "A Fast KNN Algorithm for Text Categorization," *Machine Learning and Cybernetics, 2007 International Conference on* , vol.6, no., pp.3436-3441, 19-22.

[8] T. Joachims (1998), "Text categorization with Support Vector Machines: Learning with many relevant features", European Conference on Machine Learning (ECML '98) Berlin, pp: 137–142.

[9] Sun Jian, Wang Wei, Zhong Yi-xin (2001), "Automatic Text Categorization Based on K Nearest Neighbor,"Journal of Beijing University of Posts and Telecommunications, vol. 24(1), pp.42-46.

[10] D. Koller and M. Sahami (1997), "Hierarchically Classifying Documents Using Very Few Words", International Conference on Machine Learning (ICML'97) Nashville, vol-6, pp.170–178.

[11] Thorsten Joachims (1997), "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", ICML-97.

[12] Apte C. Damerau F, Weiss S (1994), "Automated learning of decision rules for text categorization," ACM Transactions on information System, 12(3), pp. 233~251.

[13] L. Breiman, J. Friedman, R. Olshen, and C. Stone (1984), "Classification and Regression Trees", Wadsworth, Monterrey, Ca.

[14] J. He and A.H. Tan and C.L. Tan (2003), "Machine Learning Methods for Document Categorization", Journal of Applied Intelligence, vol. 18, pp. 613- 617.

[15] Kamal Nigam,Andrew McCallumzy, "A Comparison of Event Models for Naive Bayes Text Classification", School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213.