A Vocabulary, Taxonomy and Tagger for Online Distance Education

A Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of
Master of Technology

by
AMIT SHARMA

under the guidance of
Prof. T.V Prabhakar

to the
Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Kanpur, INDIA - 208016
July 2015

to Lord Chaitanya

my eternal Lord,

who propounded divine music, in the form of Sankirtana,

hare kṛṣṇa hare kṛṣṇa, kṛṣṇa kṛṣṇa hare hare
hare rāma hare rāma, rāma rāma hare hare

as the way to obtain Love of God

and revealed to us the beauty of the spiritual world,

kathā gānaṁ nāṭyaṁ gamanam api vaṁśī priya-sakhi
– *Srī Brahma-Saṁhitā 5.56*

where every word is a song, every step is a dance

and flute is the favourite companion.

Certificate

This is to certify that the work contained in the thesis entitled "**A Vocabulary, Taxonomy and Tagger for Online Distance Education**", by **Amit Sharma** (Roll No. **10327080**), has been carried out under my supervision for the partial fulfillment of B.Tech-M.Tech dual degree in the Department of Computer Science and Engineering, IIT Kanpur and this work has not been submitted elsewhere for any other degree.

Dr. T.V Prabhakar
Professor
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur
Kanpur, INDIA - 208016

July 2015

Acknowledgements

Firstly, I would like to express my heartfelt gratitude to my guide, Prof. T.V Prabhakar for his valuable guidance and support throughout my research. While he gave me a great degree of freedom in my research, his deep insights into the fundamentals of each problem inspired ideas which ultimately took the shape of this thesis. He has set before me high standards of integrity and mature vision.

I owe my sincere gratitude to my family members- my father, mother and sisters who motivate me always and set an example to do best always.

I would like to thank my friends who were always there for my help whenever I was in need. Meher Preetam helped and motivate me a lot during my thesis work. Last but not the least, I would like to give my heartfelt regards to my friend Anurag Gautam who was there for me at any point of time. Without him, this work would not have any value.

Abstract

The task of identifying keyphrases from a document has immense utility in our day to day applications which uses digital documents. Keywords - essentially - are words or phrases that are descriptive and specific to a document and can be considered the words/phrases which can best describes a document.

Keywords have become an important feature of Web 2.0 where tagging the web pages with keywords is an important way to improve the metadata set. Also they are an integral part of information retrieval systems where tasks like document classification, summarization, assessing document similarity and ranking are of immense importance. Keywords may be assigned to the document by its author. But the major anomoly in this is that it leads to poor indexing consistency. Thus one may employ professional indexers to tag the document collection uniformly. High cost incurred in hiring the professional indexers and huge time incurred in manual indexing necessitates the need of automatic keyphrase extraction.

In this thesis work, we build up a vocabulary for Online Distance Education system based on around 360 MOOC documents. We come up with a tool, used in automatic processing of the document to give all keywords in the document. Domain expert selects around 900 main keywords from these. At the top of that we build a taxonomy for these documents. Finally, we develop a software to tag a given document with appropriate tags on the basis of word frequency in the document.

Contents

List of Figures

Chapter 1

Introduction

Search for relevant information has become ubiquitous for every computer user. The most widespread use of the internet today is as an information search utility for products, travel, hobbies, and general information[1]. Although there is huge information distributed over the internet to satisfy most of the users queries, finding the relevant documents is still a considerably challenging task. This is due to the fact that the idea of making machines understand human language, although exciting, remains an elusive dream, far from reality.

Using natural language processing in addition with machine learning, statistics and domain knowledge modelling to extract keywords from the documents is feasible task and of practical use. The practice of assigning keywords to documents in order to either describe the content or to facilitate future retrieval, which is what human indexers do, has more or less been ignored by researchers in the various fields of computer science. Most systems represent document and queries as a set of keywords or keyphrases.

In this thesis, we critically examine the conventional and modern approaches to automatic keyword extraction techniques. We have build up a system for extracting the main keywords from the multiple documents iteratively. Thus we have build up our vocabulary for around 350 MOOC (Massive Open Online Courses) documents. Later we use the auto-tagging to incorporate any new given document.

1.1 Motivation

With the huge amount of digital documents existing on the internet and their growing panoply everyday, keyphrases prove to be an important metadata. Since keyphrases succinctly and accurately describe the content of the document and support thematic access to the documents, they possess added expedience in document indexing as compared to other metadata, e.g- title, author, etc.
Keyphrases can be assigned by the documents author at the time of its creation. But this yields in poor indexing consistency over the entire document collection. Thus, in digital libraries (as well as physical libraries too), professional indexers are employed to organize the documents - over the entire holding - based on their content and, to tag - with appropriate keyphrases - and finally categorize them. The keyphrases most commonly originate from a pre-defined controlled vocabulary[2]. In the absence of a controlled vocabulary the keyphrases are freely chosen important words from the document verbatim. The manual process of tagging the documents with keyphrases is labor-intensive and time-consuming and moreso it is proving infeasible in view of growing panoply of digital documents. The yearning need of automatic keyphrase extraction is evident in digital libraries.[3]

On the web, the attempt to address web search using tag-based indexed structure of the repositories is fast taking pace. Folksonomy based keyword-tagging[4] [5] wherein the person viewing/reading the web-content tags it with his own preferred tags, although useful, has several limitations including very poor indexing consistency and thus leading to unstructured document collection. Folksonomy based tagging is popular in domains like blogs and file sharing, but the pressing need of keyword-tags is equally felt in many other areas. Thus automatic keyword extraction systems that can extract and suggest a handful of keywords are believed to improve the effeciency and consistency of folksonomy based keyword-tagging.

Apart from the above, keyphrases have a variety of indirect uses, owing to their multi-functionality. Research has shown that keyphrases play a crucial role in improving the performance of various NLP applications such as automatic text clustering and classification[6], content-based retrieval and topic search [7], search result representation [8], navigation[9], automatic text summarization[10] and thesaurus construction. Thus in natural language processing, automatically assigned keyphrases would provide a highly informative semantic dimesnsion in document representation that would benefit new application. The quality of assigned keyphrases, to much extent, determine the effectiveness and utility of these approaches. Thus there is a great demand of accurate and effective methods for automatic keyphrase extraction.

Conventional approaches to automatic extraction of keywords rely on statistical calculations that involves analysis of mainly, the frequency of words occuring in the document collection. The quality of keyphrases have been improved considerably by incorporating linguistic knowledge while extracting and filtering the character strings from the documents. The work of Hulth[8] and Paice and Black[11] prove the efficacy of this method.

The idea of making the machines learn from a training corpus containing manually tagged documents to improve the keyword extraction results have been explored by Briemann[12][13]. Researchers have also used Genetic Algorithms for keyphrase extraction in whcih a set of parametreized heuristic rules have been fine-tuned to optimize the number of correctly identified keyphrases. One such system by Turney, GenEx [14] claims to give better performance than machine learning based bagged decision trees. But training GenEx on a new document collection is expensive beacuse of the high complexity of genetic algorithm involved.

Frank , Wittel and Paynter devised a novel method for automatic keyphrase extraction, KEA[15] wherein they employ the well known Naive Bayes machine learning algorithm for training over sample manually tagged documents to build a supervised training model in order to use it to choose keyphrases from the document text verbatim. The support for the use of controlled vocabulary, though added later, still could not make the system perform up to the mark in terms of indexing consistency and precision in absence of large enough training samples.

To alleviate the need of large training samples, we present the approach to learning at the knowledge model itself. The enhanced knowledge model targeted at a domain can be put to use in practical task of keyword extraction as candidate filter.

1.2 Background

As the thesis comprises of various scientific disciplines including computer science,linguistics and cognitive science, some background needs to be given about the basic interdisciplinary concepts used herein. In this section, the main concepts used in this thesis, are described.

1.2.1 Stemming and Term Conflation :

In automatic processing of natural languages one of the major issues is the variation of terms. Inflectional affixes, alternative spellings, spelling errors and abbreviation forms are, to name a few, such common variations. A concept can occur st several places in a document in different morphologically related words. While such morphologically related terms can easily be identified by a human,for a machine to recognize this, an algorithm needs pre-processing to conflate morphologically related terms to same surface form. As for example, the terms:

EDUCATE
EDUCATED
EDUCATING

EDUCATION

EDUCATIONS

can all be conflated to the common term **EDUCATE**.

Linguistic variations of words are classified into three categories namely,

(a) **syntactic variants** where the syntactic structure of the term is modified but there is no explicit morphological change,

(b) **morphosyntactic variants** which refer to cases where both syntactic and morphological changes occur over the variant and

(c) **semantic variants** where the variants are conceptually related, using synonyms and hypernyms.

Term conflation algorithms typically comprise of either morphological analyser[16] or transformational parser[17] or combination of both.

1.2.2 Vector Space Representation :

Natural language processing algorithms used to solve tasks such as keyword extraction, text clustering and text categorization necessitates explicit and unabiguous representations of the documents they need to deal with.

In NLP documents are usually representated as extensional vector. In this representation, the dimensions of the document vector are the words appearing in the document. The most universally adopted approach to term representation is termed as the bag-of-words approach: a document Δ_j is represented as a vector of term weights: $\Delta_j = (\alpha_{1j}, ..., \alpha_{nj})$, where n is the total number of elements in the set of words encompassing the document corpus and $0 \leq \alpha_{kj} \leq 1$ represents the contribution of term tau_k to the specification of the semantics of Δ_j. As far as keyword extraction is concerned, to representation and processing of documents using vector space model, three stages are required. In the first stage individual terms in a document are identified after ignoring all tokens that are stopwords- words that are very frequent and do not bear any content, e.g.- articles, conjunctions, adverbs, etc. The documents in the corpus can then be represented as a vector, with each element indicating the presence(or absence) of a term in a document. In the next stage is the terms are assigned corresponding *term weights*. It is the choice of schme of term-weigting that largely determines the efficacy of a vector space model. In principle, typically term weight, a_{ij} have three components.

$$a_{ij} = g_i \star t_{ij} \star d_j$$

Here g_i is the global weight of the i^{th} term, t_{ij} is the local weight of the i^{th} term in the jth document, d_j is the normalization factor for the j^{th} document.

1.2.3 Learning Schemes :

The field of machine learning in computer science is concerned with the design and development of algorithms that allow computers to change behavior based on the irregularities in the input data. Learning schemes are devised to interpret the changes in data patterns. The first step involves manual analysis of the input data to identify its chracteristic features. Then, the characteristic features obtained in the first step, are used to generate a learning model according to a suitable pre-defined learning scheme. In this thesis, we use supervised learning based on the Naive Bayes scheme. Although a simple method, in practice it has been proven to be effective.

1.2.4 Knowledge Models :

A knowledge model is the structural representation of knowledge by using symbols to represent pieces of knowledge and relationships between them. The term knowledge model covers the full range of tools to represent knowledge.

At a minimum, a knowledge model is a restricted list of stemmed words or terms relevant to the domain of interest that can be used for indexing. This is called controlled vocabulary. The list of terms in controlled vocabulary could grow but only under defined, guided and mutually agreed policies. Most controlled vocabularies have additional feature of see-type cross references that point from non-preferred labels to preferred label. Controlled Vocabularies are useful for providing consistency in indexing, tagging and categorizing across several indexers.

The more structured form of controlled vocabulary is taxonomy where the terms or say concepts have two kinds of relationships amongst them, namely BroaderTerm(BT) and NarrowerTerm(NT). E.g. Say for agriculture domain knowledge we have a taxonomy where there are concepts(terms), among others, Crop and Rice. Then Crop NT Rice and Rice BT Crop could be two relationships in the taxonomy. Yet more refined structured form of knowledge model is thesaurus. Thesaurus provides related terms, broader terms, nar rower terms and usedFor terms for every term present in the knowledge model. In other words, a thesaurus must clearly specify which terms can be used as synonyms(usedForTerms), which are more specific(narrower terms), which are broader terms and which are related terms. International (as well as national) standards have been developed to provide guidance on creation of thesauri. ISO 2788: 1986[19], ANSI/NISO Z39[16]: 19-2003, ISO 5964: 1985[14] and BS 8723 are, to name a few, such standards. A thesaurus also describes scope notes to clarify usage of some or all terms. The greater detail and information contained within a thesaurus compared with simple controlled vocabulary aids both the user and publisher in finding the most appropriate term more easily than in a simple unstructured controlled vocabulary. The hierarchical browsable display of a thesaurus is especially useful for a relatively large controlled vocabulary.

Ontology is the climax of a contolled vocabulary where the concepts or terms have no restriction in the relationships they could have with each other. Concept Map[4] is one excellent example of ontology.

1.2.5 Evaluation Parameters :

The standard IR parameters Precision and Recall is based on the # of matchings between two tag-sets. Let us consider two tag-sets α and β Then precision of βas compared to α is defined as:

1.3 Thesis Outline

In this introductory chapter, we stated the research problem and gave the motivation and background of this work. In the next chapters, we will investigate the congnitive models of keyword extraction, explore the existing approaches to keyword extraction, describe building-up of the vocabulary, applying KEA algorithm for any new document and to incorporate the new keywords in the vocabulary. The domain of documents which we have taken is the online distance learning documents or MOOC documents.

Chapter 2 discusses about cognitive models of keyword extraction. of the content of a document. A model of text comprehension by indexers (including classifiers and abstractors) is presented. In this chapter we dive into the history of keyword extraction and indexing. The relationship between text and index entries is examined and progress towards the devlopment of cognitive process model is studied. This chapter also discusses some important types of indexes and concordances, in vogue.

Chapter 3 explores the existing approaches to automatic keyword extraction. In this chapter we present a glimpse of large number of existing methods for automatic indexing published in the literature. Also we look on the well known current implementations for automatic keyword extraction.

In chapter 4 we present methods of building up the MOOC vocabulary. We are given around 360 Online Distance Education documents and a domain experts selects the main keywords from all possible words. This chapter describes how to make set of all possible words from the set of MOOC documents.

Chapter 5 discusses about the making of Taxonomy in a given document and input vocabulary. The making of MOOC-tags from MOOC-vocab also has been described. Finally we tag a input document on the basis of frequency with the help of these tags.

Chapter 6 concludes the thesis with a brief glimpse into the future work and scope for further improvements.

Chapter 2

Cognitive Models of Keyword Indexing

2.1 History

The task of keyword extraction, indexing and abstracting have evolved in somewhat natural way, dating back to second millenium B.C. when scholars developed the custom of abstracting the books. The first systematic approach to indexing emerged in fourteenth century. It was true alphabetical indexing. The indexes of this era consisted primarily of the keywords in the theses, or desputations, alphabetically arranged.

Then, in the fifteenth century, the first example of catalogue with a subject index turned up. The catalogue then had three indexes - author, subject categories and anonymous works listed by the catchwords in their titles[18]. It was in the seventeenth century that the first recorded catalogues with subject indexes were produced. During this time, Adrien Baillet,[18] a librarian, prepared a catalogue of Lamoignons library, with the main arrangement in classified order, and an alphabetical subject index. In the practice to his catalogue, he described the system: it included cross references, specific entry, subdivisions, and rules for choosing aong alternative words and phrases for subject headings.

Subject indexes were in vogue by eighteenth century. But the choice of terms and order of entries in the index remained haphazard. Later with the development of taxonomy in library work and documentation, subject indexing became more systematic.

Until the beginning of twentieth century except for implicit facet analysis in Dewey Decimal classification[20] and Kaisers concrete-process system[19], there were no major advances in indexing methodology. It was the onset of world war II that brought about the need to provide a key to the growing mass of information. This period witnessed a change in the attitude of publishers, librarians, and information users. Techniques were developed for indexing more quickly, consistently and thoroughly with the noticeable reault of the invention and popularization of post-coordinate indexing by Taube, Batten and others. Post-coordinate indexing[21] involves breaking up the complex concepts into its components for indexing. The seracher then can search on any of the parts in any order.

The next landmark in the development of indexing techniqiue was the advent of greatly extended form of authority tool called thesaurus in late 1950s. It grew out of old subject heading lists, but is more carefully compiled usually with a fully worked-out structure, in contrast to subject heading lists. In 1960s the idea of coextensiveness was formally developed by Coates as a new solution to the problems of specific entry. Coates proposed a new definition of specificity: the specific heading is the one that is coextensive, summarizing all aspects of the subject of a work. This gave the seed for the development of PRECIS(Prserved Context Indexing System) developed by Derek Austin for British National Bibliography[25].

The use of computers and media other than paper to store indexes was witnessed in late 1950s. This period also brought forth efforts to evaluate indexing using experimental and quantitative methods. The Cranfield Project by C.W.Cleverdon[22] was the first historical experiment on evaluation of indexing systems. Here Cleverdon made a comparison of four index languages using a large data base and some 1200 searching subjects. It was the first significant use of recall and precision measures to evaluate the performance of indexing methods.

Next emerged the development of KWIC(key word in context) indexing by Luhn and his colleagues at IBM[23]. They did not invent a derivative indexing; it preceded assignement indexing by centuries. Yet it was a major intellectual advance, in the sense that its contribution was to find a way to do the indexing fast and cheaply and with no intellectual input to the individual document analysis.

After this, the period of 1960s and later witnessed a great deal of experimentation and research into various forms of automatic document surrogation. Salton and Sparck Jones were the leading researchers[11] [17] [23]. Computers began to be used as data processing machines in the preparation of manuscripts since publication has made it possible to begin preparation of indexes at an earlier stage in the publication cycle to update them more regularly and to search them more conveniently.

As long as indexing was an operation necessarily performed by people, word indexing was generally confined to personal and other proper names. While subject terms were usually created by both derivation and assigment. But with advent of computer, it became possible to create subject indexes by mechanized word indexing. With phenomenal progress made in this respect, there are strong indications towards the refinement and popularization of automatic keyword extraction and indexing.

2.2 Nature and Type of Indexes

This section deals with the generic indexing rules immaterial of the type of document indexed and the method of indexing. Specifically we delve into the criteria used by the librarians for the selection of terms to be used as keyphrases and titles in displayed indexes, as descriptors in non-displayed indexes, and in the vocabulary management component of indexes. The synonymous, equivalent, hierarchical, and associative relationships among concepts of the documents play a significant role in the choice of terms for indexing.

2.2.1 Function of an index

The intended purpose of an index, essentially, is to give an effecient means, to users, for locating relevant documents with the information requests. The usefulness of indexes are listed as follows[22]:

- ∗ identifying documents or documentary units dealing with specific topics and/or possessing specific features.
- ∗ indicating all crucial topics or features of documents according to the degree of comprehensiveness of the index.
- ∗ providing access to topics or features customized for prospective users.
- ∗ using appropriate terminology that is as specific as documents authorize.
- ∗ providing access through synonyms.

2.2.2 Types of indexes

Although all indexing practices have the common underlying philosophical base, that is, to provide a guideline for the users to the intellectual content and physical location of documents, for practical purposes there are several levels or strata of indexing forms, approaches and uses. Indexes may be classified on the basis of the type of object referred by the headings; by nature of term used for index headings; by method employed in arranging entries; by methods employed in term coordination, etc. Some of the commonly used indexes used by librarians are enumerated as follows:

1. **Name Indexes**

   A name index is an index to the names of people cited or referred to in a document. Such name indexes are, in effect, subjects and can become headings in a subject index.

2. **Author Indexes**

   An author index is an index to the author of various items in the work being indexed . Their entry points are people, organization, corporate author, government agencies, name of universities, etc. . Author indexes are basically used to guide searchers to subjects rather than the names of authors. An author index in conjunction with subject, concept or topic index brings about a one- alphabetic product that is simply called index. Research has given clue that in certain cases, authors are strong indicators of subject-content in a group of documents.

3. **Alphabetic Subject Indexes**

   The term alphabetical index covers a number of different kind of indexes. Though the arrangement mostly observed, is alphabetical order, it is not the only method that can be used. The index may follow a classified arrangement, or both alphabetical and classified at the same time. Generally any classified index will need an alphabetical subject approach to supplement it, either separate or built-in form- to make the use convenient and effecient.

   The subject index, also called topic index or feature index takes into account of the topics treated in the documents and/or features of documentary units. An alphabetical index is based on the orderly principles of letters of the alphabet and is used for the arrangement of subheadings, cross-references, and qualifying terms, as well as main headings. All entry items are in alphabetical order, including subject terms, author names and place names.

   The major advantage is that it follows an order familiar to us. While the major drwabacks with the alphabetical arrangement are the problems of synonymity and scattering of similar entries. Scattering refers to that situation where the subcategories of a subject are not drawn together under the generic term but are dispersed throughout the list. The technique to overcome this problem is to use the syndetic system, i.e. the frequent use of cross-reference from the unused term to the preferred term. The see-also cross-reference is used most often to bring attention to related topics or headings as genus is to species or the reverse.

4. **Classified Indexes**

   Classified indexes refer to the indexing scheme where entries are arranged in a hierarchy of related topics, starting with generic topics and working down to the specific. Combination of alphabetical and classified indexes are possible. In practice, the major headings are arranged in a classified order; the entries under them can be in alphabetical order. Alternatively the major headings can be alphabetized and the subheadings or entries under them can be placed in a classified order.

   Classified indexes have some advantages and disadvantages. Because of their conceptual approach, they aid significantly in search. The subject arrangements attempts to pull together related subjects. In other words, the like entities (concepts, topics or subjects) are brought together whereas alphabetical order tends to scatter related terms. Classified indexes make searching simple if the user wants to conduct generic searches, since the hierarchy is visually presented. When an entry is located, the user is immediately made aware of the closely related concepts or items. It is tantamount to browse a library stock.

   The major disadvantage of a classified index is that it necessitates a secondary file, an alphabetical list to help locate the right position in the classified list. An alphabetical guide to the classification scheme, which is called an index to the classification scheme is an example. This added alphabetical guide has been found to be necessary because what may be logical to the classified may appear unexpected for the user.

5. **Subject Indexes**

   The term subject index is used to refer to all the indexes that are not clearly identifiable as author, title or special indexes. A subject index is also different from topic, concept and word indexes. Subject are the foci of a work the central themses towards which the attention and efforts of the author have been directed. Subject indexes are specifically useful for those searchers who know the subject that they are interested in, and want to browse lists of documents on that subject.

Process of subject indexing : In the process of subject indexing, subject indexers first identify subjects to be indexed. Secondly he has to rmbody the subject in words to paraphrase the subject. Subject paraphrases are as complete as necessary, terse and independent of other contexts. That is they do not omit significant parts of the subject studied; they are not redundant. They dont have pronouns, verbs, adverbs or articles. For the profesional indexer, the subject paraphrase is not written, but coined and carried in memory during the indexing. The specificity of the subject paraphrase is made great enough to enable the index entries derived from it to be distinguished from all other subject entries from other documents that appear under the same subject headings in the index.
Next subject indexers create the subject index entries from the paraphrases. Each entry consists of the subject heading, a modification(modifying phrase) and a refernce in that order.

Term Alteration : Once words or terms that are useful as guides to the subject paraphrase have been selected by the subject indexer, the next step is to alter the few terms that need translation into standard index headings. Subject index headings are standardized to eliminate scattering of like entries. Quite often, most subject headings will be identical with the words used by the authors. Much of the standardization of index heaidngs involves the singular and plural forms of the words used. It also involves the conversion of adjectives and gerunds into nouns.

Modifications : Once a subject heading term has been selected and translated into a standard subject heading, the next step is for the subject indexer to coin for the modification or to select a subheading. Modifications make the index entry much more specific than it would be without them. Modifications are coined to be as specific as necessary to differentiate among all entries having the same heading in the index. Such a great specificity enables the searcher to decide definitely for or against looking up the entry. Subheadings are akind of modification that are used in place of coined modifications. Subheadings are selected from a list of standard subheadings. Modifications are coined ad-hoc by the indexer to fit the subject studied.

6. **Word Indexes and Concordances**

   Word and name indexes, which are sometimes called concordances, are indexes to individual names and words that are author used and in one sense most closely represent the information and ideas the author had in mind when writing. Although the thoughts of an author are expressed in words, but not all the words lead to the subjects, concepts or topics presented in a work. An index to all the words in a document is a concordance or word index.

   With the advent of computers and techniques of language data processing , many new word indexes have been created, such as KWIC, KWOC and permuterms. It is to be noted here that word indexes are the most bulky, concept indexes are the next most buly, topic indexes are the next most and subject indexes the least bulky.

Distinction between subject indexes and word indexes: For guidance to subjects, subject indexes are much more effecient than many other indexes. Users of subject indexes do not waste time consulting concepts, topics and words in which they have no interest. Such indexes are different in scope and function from subject indexes. Also subject indexes are selective and the terminology is controlled so as to minimize scattering. Word indexes are not selective nor is the vocabulary controlled. But they have their own unique qualities and uses.

KWIC, KWOC and Concordances : Here we dwelve into two types of word indexes - those derived from the titles of the articles and those derived from the full text. The first category includes KWIC indexes, KWOC indexes and permuted indexes. The second category indexes are commonly called concordances.

KWIC : KWIC is an acronym for Keyword in Context. It is based on three principles : (a) that titles are generally informative
(b) that words extracted from the title can be used effectively to guide the searcher to an article or a paper likely to contain desired information
(c) that although the meaning of an individual word viewed in isolation may be ambiguous or too general, the context surrounding the word helps to define and explain its meaning.

In the KWIC format, the natural word order of the title or other text segment is preserved on both sides of a keyword; the keyword is arranged in alphabetical order down the center of the page or column with surrounding title or text segment to the right and left. The major advantage of KWIC indexes is that they are easy to build using computer application[28]. But the drawback is that it requires special skill to search effectively by users.

KWOC : Since the filling word in KWIC indexing is not in the biginning place, a further development was made to move the keyword back to its normal place ath beginning of the line, but to follow it by the complete title, rather than by some altered form as in catchword indexing[64][28]. This is called Keyword out of Context or KWOC. The KWOC format preserves the traditional format of an index with the lead keyword on the left, followed by the title or other text segment as a subheading.

Permuterms : Permuterms or permuted indexes display every possible combination of index terms or descriptors. Such terms are selected by indexers or extracted from text according to various computer algorithms. Here we have heading-subheading combination for each word pair. In this system the actual titles are not displayed, but the words which form the basis of the index entries, are extracted from the titles. Pairs of keywords are extracted from each title to be indexed.

Concordances : Concordance is an alphabetic index to all the words in a single text. Each word that is present in the text is an index entry. This type of indexing involves the selection of words, often without discrimination from running text, and their display in context. For practical purposes, the entries in the concordance are limited to the principal words in the document.

Chapter 3

Existing Approaches to Automatic Keyword Extraction

The previous chapter threw some light on the history of indexing, abstraction and keyword extraction. Various cognitive models of indexing have been discussed along with their relevance to appropriate applications. The various models studied propounded various strategies which depended on whether the keywords got selected from texts in the document or terms given to similar type of documents or whether they came from a closed set of words.
The primary focus of this chapter would be to explore existing approaches in the area of automatic extraction and assignment of keywords. A couple of approaches will be discussed

3.1 Keyword Assignment from a Vocabulary

Keyword assignment is generally done from a pre-exisitng set of potential keywords. Generally, two different ways are followed:

1. **Text Categorization :** Text in the document as words and phrases are analysed and the document is segregated into various categories listed in the vocabulary using automatic or manually generated rules.

2. **Candidate Keyword Generation :** It is to be noted that in both of the above techniques, the features of the dcoument are analysed rather than the words to determine the keywords from the prespecified vocabulary set.

In either of the two approaches, the characteristics of the documents, rather than the words verbatim occuring in them, are analyzed to determine keywords from the controlled vocabulary.

3.1.1 Text Categorization: Classification and Rule Binding

Throughout a span of fifty years, a large number of techniques have come into being throught the efforts of many library scientists[60][39][50]. Initially, knowledge-engineering experts manually created the rules for classification and applied to electronic documents. Fuhr and Knorz (1984)[54] devised a decision-making system that employs more than 150,000 such rules to map physics documents to vocabulary terms. The rules have the form

IF ⟨ property⟩ is identified in the document then DESCRIPTOR.

where ⟨ property⟩ is a word, phrase, or a physics formula that appears in the text.

As time progressed, machine learning gained precedence in research circles[60]. Automatic classification rules were generated through careful application of learning schemes over manually classified document sets. A confidence value between 0 and 1 is assigned to every rule and an algorithm assigns documents to a particular class based on a threshold value.

In the vector space model for text classification, the documents take the form of vectors with each word appearing in the collection representing a dimension. The element in the vector marks the weightage of a particular word while also indicating its presence or absence in a particular document. A classifier analyzes the similarity of a new document to the oens which are manually assigned to vocabulary terms and finds out relevant keywords.

Inundating the vector space with large number of words increases the dimension of the vector space entailing significant computation. To reduce the number of words various techniques have been employed in literature including stemming, ignoring stop words and retainment of words determined computationally with an appropriate function[60]. Dumais et al. (1998)[61] compared the effectiveness of text categorization performance of classifiers using more sophisticated techniques such as considering mutli-word index terms, using shallow parsing and deep parsing[44].It was observed that the accuracy of these classifying functions couldn't be improved through linguistic techniques[51].

There has been a plethoric use of learning algorithms for the purpose of text classification. The performance of classifiers like Find Similar, Decision Trees, Support Vector Machines etc., have been studied by Dumais etal.(1998)[61]. The SVM method of classification was found to achieve the highest accuracy of 87for classifying news stories into 118 Reuters categories. This metod can handle more dimensions without setting the parameters.

Similarly, Sebastiani[60] applied classifier committees, where the decision rests on a collection of classifiers that can be combined in multifarious ways. The decision with the majority vote among the collection of classifiers will be finally chosen.

There are two major concerns in vector-based classification, one being that every vocabulary term needs a separate classifier while the other being the computational resources required when the problem turns many dimensional. The former problem has been dealt by Pouliquen et al.[56] by using a large training corpus. The latter problem is essentially dealt by limiting the analysis to most common words[62].

Plaunt and Norgard(1998)[27] came up with the technique of generating association rules using a contingency table to tackle the problems mentioned afore. This table mentions the number of co-occurences of a phrase with manually assigned vocabulary terms. A confidence value is computed between a phrase and a vocabulary term using likelihood ratio statistic over co-occurrence values. Precision and recall values were found to be 21% and 64%.

Another work conducted by Aronson et al. (2000)[23] and Mark et al.(2003)[53]

applied rules governed by conditional probabilities of vocabulary terms to simultaneously occur with document phrases in training corpus. Large corpus has been considered to ensure high coverage of vocabulary. Aronson et al. broke words into bigrams and trigrams of characters while Mark et al. broke them into subwords. The latter approach is based on a manually created dictionary that ensures orthographic, morphologic and semantic normalization of document terms into a group of identifiers. The individual probability of a vocabulary term being a keyword is found by multiplying conditional probabilities of the subword trigrams and dividing by probability of trigrams present over all the documents under training. The precision and recall woud found to be 30%.

Text classification based methods provide scope for terms not exactly present in the document to be assigned as keywords since a classifier is a generalized model of representing the keyword through phrases, words, etc. These classifiers born out of machine learning techniques give accurate results when small vocabularies are used[61] or huge traning data is used[56]. These should be considered valid problems of text classification.

3.1.2 Candidate Keyword Generation and Filtering

This method is particularly helpful in cases where the controlled vocabulary is huge. In the beginning, candidate topics are found by mapping phrases in the document to vocabulary terms. Further, significant candidates are computed based on their properties. Training data need not be given for each vocabulary term. But, this technique broaches up new problems. Mapping has inherent relationship with language phenomena like synonymy and polysemy. In the filtering pcocess,certain properties which can distinguish topics from non-topics need to be found.

3.2 Keyword Extraction from Text

In this approach the goal is to extract keywords from the text without assignment of any pre-defined vocabulary. This is carried out in the following two steps:

1. **Candidate Keyword Generation :**

   In the first step, all content bearing words and phrases are extracted from the document by getting rid of stopwords. The phrases generally consist of combination of content bearing words along with stopwords. In other scenarios, use of a PoS tagger and a shallow parser can lead to extraction of NP phrases, under the assumption that keywords are usually nouns. The computation of candidate topics fis usually a step-by-step process. This is done by extracting sequences of words of known lengh called n-grams.These n-grams, are then matched against terms present in the vocabulary. Before matching, the sequences and vocabulary term need to be free from stop words. If the vocabulary contains non-descriptors, document phrases are matched against them to determine the corresponding descriptor. This process is called semantic conflation. This is the major advantage of using controlled vocabularies: terminology present in a whole lot of documents can be reduced to a set of controlled terms. If more than one vocabulary match is possible, word sense disambiguation is required.

2. **Filtering** In the second step, filtering, important properties that can separate keywords from non-keywords need to be found. Analysis is carried out for selected candidates in regard to a specific feature which can show its nearness to a potential keyword. Some additional filtering heuristics have also been tried for this. The description of these techniques is given in the following subsections.

3.2.1 Machine Learning based Filtering

Keyword extraction and filtering can be defined as a machine learning problem in which sample documents and their resultant keywords can be given as training data set. Once this is learnt, keywords can be extracted out of unkonown documents. The key component in this learning process are prominent features of the information which can help differentiate between a keyword and a non-keyword. Such features need to be calculable for all the wordd or group of words present in the document. Following are certain prime implementation of machine learning techniques applied to keyword extraction:

1. **GenEx** P.Turney[55] is the creator of GenEx algorithm. He initially applied a general purpose learning algorithm C4.5 for extracting keywords. He later found that a tailo- made algorithm enhances precision and thus shifted to Genertic Algorithm. The word GenEx refers to Genitor[67] and Extractor[55]. Genitor is a genetic algorithm whereas Extractor is meant for keyword extraction.Genitor tunes various parameters of Extractor so that the performance of Extractor is optimized. The Extractor in turn uses a set of heuristic rules to generate a list of keywords.

   Figure 3.1: Genex versus C4.5 (taken from [55])

   The rules present in extractor are determined by twelve parameters. These parameters are decided by a supervised machine learining algorithm. The algorithm is tuned with a dataset, consisting of documents paired with target lists of keyphrases. The comparison between the performance of GenEx with C4.5 algorithm and the results as given by Turney can be found in the figure 3.1.

2. **Hulth** The speciality of Hulth lies in the additional use of NLP techniques. Her work consisted of examining multiple methods to select candidate phrases

like NP based chunking, PoS tag sequence matching and n-gram based selection. The best precision evaluated using NP-chunks is 29.7 % and the best recall percentage is 66 when POS patterns are incorporated. Certain features used during the filtering stage were: PoS-tag, term frequency, position of first occurence and inverse document fequency. A combination of different suitable learning algorithms along with NP based selected candidate phrases got best results.

3. **KEA** Frank et al.[42] used simple and robust methods to develop Keyphrase Extraction Algorithm, KEA. The process requires to extract n-grams, apply iterated Lovins stemming, and filtering. The filtering stage necessitates computing of two features for all of the selected phrases which are TF X IDF and position of first occurence. While training, the Naive Bayes learning algorithm creates a model from manually tagged documents. During filtering, for each candidate, an overall probability is calculated from the model to check if it can be a keyphrase . A rank is given to each of the candidate phrases on the basis of probability amd k number of top ranked phrases are deemed to be keyphrases, where k is given by the user.

3.2.2 Symbolic Heuristic Methods

In these methods, a manual approach is taken to analyze documents and their keywords and come up with certain filtering heuristics.Two popular techniques in this domain are as follows:

1. **Paice and Black :**Initially, n-grams get extracted which undergo transformation to pseudo-phrase.The stop words are removed from all n-grams and then stemming takes place upon filtered phrases. A final alphabetical sorting takes place. A score is computed for each pseudo-phrase using a novel formula:

$$score = W * (F - 1) * N^2$$

Here W refers to the sum of the weights of all words in the pseudo-phrase, F refers to the frequency of the phrase in the document and N refers to phrase length in words.

A pattern- based technique is then employed to find the relations between selected candidates. These patterns are independent across domains. The phrases that can match these patterns are considered to be output of this algorithm.

2. **B & C :** Barker and Cornacchia[48] developed this tool. The candidate generation phase is taken care of using simple dictionary lookup. The greatest advantage of this is that the online dictionaries list the root form of each word and allow such phrases to be treated as good schema. At the end of candidate extraction, filtering is done. In this stage, frequency of each head noun which appears in the document as the head of some noun-phrase is computed.The best N heads get chosen and the scores of all these phrases are computed. The best r noun phrases are taken to be keyphrases. The values of N and r are set heuristically or on an empiric basis. Barker and Cornacchia also found the performance of their tool is at par with Turneys extractor[55].

3.3 Combining Extraction and Assignment

The preceding sections discussed keyword extraction and keyword assignemnt. It is evident that both of them have their own set of pros and cons. Keyword extraction appears to do well even without controlled vocabulary, but it is necessary to have a sufficiently large training corpus. The performance is also not comparable to manual indexing. Further, it has been found that sometimes extracted keyphrases are useless and non-grammatical due to inherent errors imbibed within linguistic techniques.

In the case of keyphrase assignment, controlled vocabularies are necessary for specific domains. Inspite of tyring to restric the domain to encompass narrow field, the controlled vocabularies happen to contain several thousands of terms[15].

3.3.1 KEA++

Medelyan and Witten tried to get the best of both worlds(extraction and assignment) by adding both their advantages and removing both their disadvantages. They enhanced the existing KEA algorithm by adding value to the candidate identification process. Here, terms from a controlled vocabulary that are available in the document are taken to be candidate phrases. Once the stopwords get removed, only those content bearing words are taken which are present in the controlled vocabulary. The learning process remains as usual. The enhanced KEA algorithm is discussed as follows:

1. **Candidate Selection and Term Conflation :**

   The document is broken down into tokens with the help of white spaces and punctuations. In the process, any word consisting of numerals get removed. The n-grams undergo to form pseudo-phrases[45]. These pseudo-phrases are matched against conflated vocabulary terms.

2. **Keyphraseness Features**

   Various features have been defined that can indicate the keyphraseness of a candidate term.

   (a) **TF × IDF**

   TF × IDF is found as follows:

   $TF \times IDF(\pi, \delta) = \frac{freq(\pi,\delta)}{size(\delta)} \times -log_2 \frac{docf(\pi)}{N}$

   In this context, $freq(\pi, \delta)$ represents the number of occurences of π in δ, size(δ) represents the number of words in δ, docf (π) refers to the number of documents(in the global collection) containing π, and N is the total number of documents in the overall global collection.

   (b) **Position of first occurence, f :**

$f = \frac{noofwordsoccuringbeforethefirstoccurenceofphraseinthedocument}{noofwordsinthedocument}$

(c) **Length of the candidate phrase :**

This feature is a booster used especially in the case of phrases containing two words.

3. **Learning**

   For every document, pseudo phrases are selected from the training corpus followed by computation of feature values for every individual.The pseudo phrases gets segregated into positive and negative instances depending on whether they were manually given or not. Naive Bayes algorithm implentation of Weka[34] is used to generate the prediction model for this purpose.

4. **Outputting Keyphrases** The probability of candidate phrases from new documents are considered for computing their keyphrasedness out of the model previously generated from the training phase. Initially, all the four feature values are computed for all candidate phrases, say $TF \times IDF$ comes out to be t and distance of first occurence is f , length is l. Then two new quantities, $P[yes]$ and $P[no]$ are calculated as described as follows.

   $$P[yes] = \frac{Y}{Y+N} P_{TF \times IDF}[t|yes] P_{position}[f|yes] P_{length}[l|yes]$$
   $$P[no] = \frac{N}{Y+N} P_{TF \times IDF}[t|no] P_{position}[f|no] P_{length}[l|no]$$

   Here Y is the number of positive instances in the training corpus and N means the number of negative examples. $P_{TF \times IDF}[t|yes]$ gives the conditional probability that $TF \times IDF$ is t when a particular instance is positive. The other terms refer to corresponding probability values.

   The final keypharsedness probability of a candidate term is computed as,

   $$p = \frac{P[yes]}{P[yes]+P[no]}$$

   The top r, (as requested by user), candidate phrases having highest probability given out as keywords.

Chapter 4

Building-up MOOC Vocabulary

Building up controlled vocabulary or Thesaurus has immense importance in the context of data mining and in the field of Natural Language Processing as well. Selecting the words of interest or Positive words from the data and given documents is always crucial. Domain experts select these terms in order to enrich the vocabulary properly.

4.1 Purpose of Vocabulary Control

Vocabulary plays important role in Information Retrieval Systems (IRS)

The term information retrieval is generally used to refer to the activities involved in locating documents (books, periodicals, articles, reports, and other forms) dealing with particular subject matter, and an information retrieval system consists of a group of activities and components designed to facilitate access to the subject matter of documents.

Descriptive cataloguing and subject cataloguing or subject indexing are applied to the items selected. Conceptually, subject cataloguing is the same as subject matter of complete publications (e.g., books, periodicals) while the latter is more likely to apply to parts of publications (e.g., periodicals articles). In this regard, the term indexing refers to the subject indexing although the principles discussed apply equally to subject cataloguing.

The objectives of vocabulary control within an Information Retrieval System (IRS) can be summarized as follows:

1. To promote consistent representation of subject matter by indexers and searchers, thereby avoiding the dispersion of related materials, through control (merging) of synonymous and nearly synonmous expressions and by distinguishing among homographs.

2. To facilitate the conduction of a comprehensive search on some topic by linking together terms whose meanings are related paradigmatically or syntagmatically.

4.2 Building vocabulary from a single document

Extracting keywords of interest from a given document is the key step in building up vocabulary. Let us say, we are given some document in the pdf format. So our first step is to convert that pdf to text file and perform some operations on it.

4.2.1 Some Terminologies :

We will give some term definitions now in order to make the things more clearer.

Tokenization :

In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis. Tokenization means to break the input text file into the set of tokens containing the file. As a result, tokenization of any file will give all the words or tokens which are present in that particular file. Let us see an example:

Stemming :

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form. The stem needs not to be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation. Stemming programs are commonly referred to as stemming algorithms or stemmers.

A stemmer for English, for example, should identify the string ”cats” (and possibly ”catlike”, ”catty” etc.) as based on the root ”cat”, and ”stems”, ”stemmer”, ”stemming”, ”stemmed” as based on ”stem”

Stopwords :

In computing, stop words are words which are filtered out before or after processing of natural language data (text).[1] Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all processing of natural language tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search.

Positive Words :

Positive words are the words of interest or the words which we want to get include in out vocabulary. There should be a domain expert to select these words.

Negative Wods : Negative words are the words which we do not want to include in our vocabulary. Later we will see in this chapter, that bag of negative words is also as important and will help us to build the vocabulary from multiple documents.

After getting the text file from the pdf document, first we will tokenize the file to get a bag of all the words. Then we perform stemming, i.e., we will get stemmed

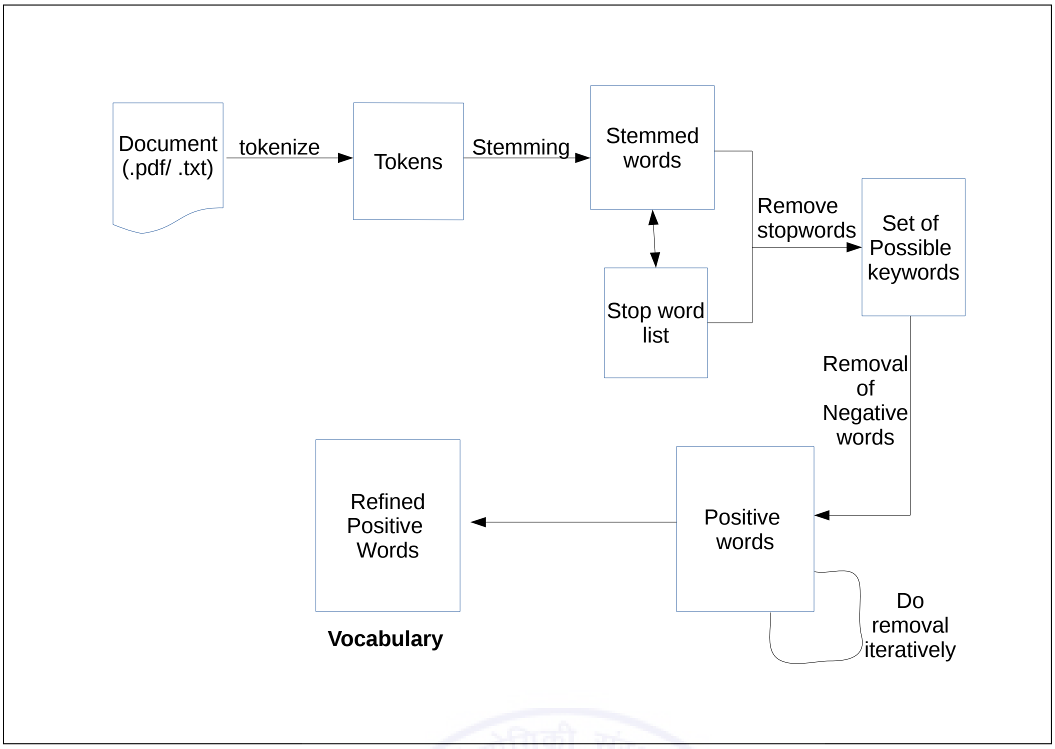


Figure 4.1: Different stages in getting the main keywords from the given input document.

words of all the tokens. Once we are done with stemming, we will remove all the stop words from these words. Then we will get the set of possible keywords.

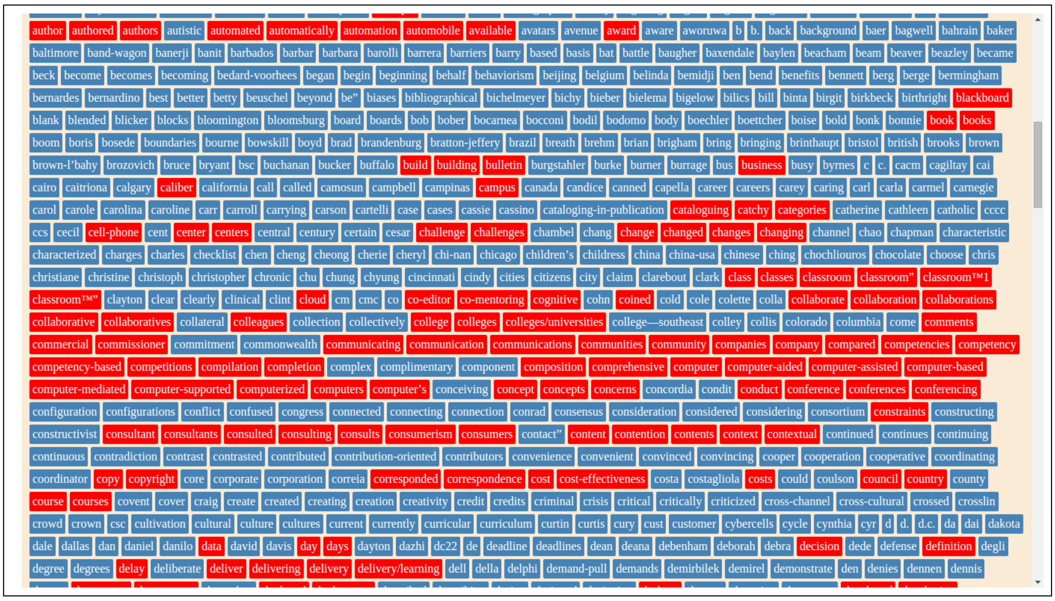
Now comes the role of domain expert in selecting the positive words of interest which can be the keywords in our final vocabulary. We will remove the negative words iteratively from this set of words. We will do many such iterations to refine our vocabulary. Finally, we will get the vocabulary.

4.2.2 Techniques Used

There have been various techniques used for building up the vocabulary from the documents in a given domain. We describe the following two main techniques to extract words from the posible set of words to build up the vocabulary.

Inclusion of Positive Words

Given the document from which we have to extract the main keywords, we have to convert it into text file. After that, we tokenize the file, remove stopwords and then



author authored authors autistic automated automatically automation automobile available avatars avenue award aware aworuwa b b. back background baer bagwell bahrain baker
baltimore band-wagon banerji banit barbados barbar barbara barolli barrera barriers barry based basis bat battle baugher baxendale baylen beacham beam beaver beazley became
beck become becomes becoming bedard-voorhees began begin beginning behalf behaviorism beijing belgium belinda bemidji ben bend benefits bennett berg berge bermingham
bernardes bernardino best better betty beuschel beyond be" biases bibliographical bichelmeyer bichy bieber bielema bigelow bilics bill binta birgit birkbeck birthright blackboard
blank blended blicker blocks bloomington bloomsburg board boards bob bober bocarnea bocconi bodil bodomo body boechler boettcher boise bold bonk bonnie book books
boom boris bosede boundaries bourne bowskill boyd brad brandenburg bratton-jeffery brazil breath brehm brian brigham bring bringing brinthaupt bristol british brooks brown
brown-l'bahy brozovich bruce bryant bsc buchanan bucker buffalo build building bulletin burgstahler burke burner burrage bus business busy byrnes c c. cacm cagiltay cai
cairo caitriona calgary caliber california call called camosun campbell campinas campus canada candice canned capella career careers carey caring carl carla carmel carnegie
carol carole carolina caroline carr carroll carrying carson cartelli case cases cassie cassino cataloging-in-publication cataloguing catchy categories catherine cathleen catholic cccc
ccs cecil cell-phone cent center centers central century certain cesar challenge challenges chambel chang change changed changes changing channel chao chapman characteristic
characterized charges charles checklist chen cheng cheong cherie cheryl chi-nan chicago children's childress china china-usa chinese ching chochliouros chocolate choose chris
christiane christine christoph christopher chronic chu chung chyung cincinnati cindy cities citizens city claim clarebout clark class classes classroom classroom" classroom™1
classroom™" clayton clear clearly clinical clint cloud cm cmc co co-editor co-mentoring cognitive cohn coined cold cole colette colla collaborate collaboration collaborations
collaborative collaboratives collateral colleagues collection collectively college colleges colleges/universities college—southeast colley collis colorado columbia come comments
commercial commissioner commitment commonwealth communicating communication communications communities community companies company compared competencies competency
competency-based competitions compilation completion complex complimentary component composition comprehensive computer computer-aided computer-assisted computer-based
computer-mediated computer-supported computerized computers computer's conceiving concept concepts concerns concordia condit conduct conference conferences conferencing
configuration configurations conflict confused congress connected connecting connection conrad consensus consideration considered considering consortium constraints constructing
constructivist consultant consultants consulted consulting consults consumerism consumers contact" content contention contents context contextual continued continues continuing
continuous contradiction contrast contrasted contributed contribution-oriented contributors convenience convenient convinced convincing cooper cooperation cooperative coordinating
coordinator copy copyright core corporate corporation correia corresponded correspondence cost cost-effectiveness costa costagliola costs could coulson council country county
course courses covent cover craig create created creating creation creativity credit credits criminal crisis critical critically criticized cross-channel cross-cultural crossed crosslin
crowd crown csc cultivation cultural culture cultures current currently curricular curriculum curtin curtis cury cust customer cybercells cycle cynthia cyr d d. d.c. da dai dakota
dale dallas dan daniel danilo data david davis day days dayton dazhi dc22 de deadline deadlines dean deana debenham deborah debra decision dede defense definition degli
degree degrees delay deliberate deliver delivering delivery delivery/learning dell della delphi demand-pull demands demirbilek demirel demonstrate den denies dennen dennis

Figure 4.2: Selection of positive words from the set of all possible keywords.

stem the resultant words. The whole procedure has been described in above section. Now we get the all possible candidate list of words which can be in our positive bag of words.

From this set of possible keywords, we select the postive words to include them into our vocabulary. Once we have selected the positive words from the set of all words, we again go through the remaining words to find out any missing positive word there. We do this iteration several time in order to get the final vocabulary for a document.

Exclusion of Negative Words

The other technique is to exclude the negative words from the all possible bag of words. That is, we will remove a word one by one in order to refine the vocabulary. We do this iteration many times in order to enrich our vocabulary. The remaining words will be in the bag of positive words.

There should be domain expert to choose these positive words or negative words in that particular domain. Although, we have used the second technique since there is chance of skipping a positive word in the first technique. So we remove the negative words iteratively to build the final vocabulary.

4.3 Combining vocabulary from multiple documents

Previously, we have seen how we can build vocabulary from a single document. Now, we have task of building vocabulary from multiple documents. We have a set of around 360 pdf documents.

Let us say, we have n different documents. We will denote it as: $d_1, d_2, d_3...d_n$; where d_1 represents the first document.

Trivial Approach :

In this approach, we will build the vocabulary for individual document and then take the union of all the vocabularies. Thus we will get the vocabulary for all the documents.

$$\text{Vocabulary} = \Sigma d_i^V$$

where d_i^V represents the vocabulary for the i^{th} document.

But we can do something good here. As we see, first we are developing vocabulary for each document. That means we have to remove negative words from each document and some of these negative words may appear in multiple documents. So we have to remove these negative words many times.

Iterative Approach :

As we see in above approach, we have to build vocabulary for each document first and then take summation. There may be some negative terms in all the documents and we have to remove them from each of the document.

In this approach, we keep track of negative words also from each document. We subtract these negative terms in a document d_i from the total words in document d_{i+1}. The resulting token would be the reduced set of words. Now the domain expert would select positive terms from this reduced set of words. We do it iteratively for all the documents and hence, finally we will have all the postive words from all the documents together.

For any i^{th} document, we can write:

$$d_i = d_i^+ + d_i^-$$

where;

d_i represents the total bag of words in the i^{th} document;

d_i^+ denotes the postive bag of words in the i^{th} document;

d_i^- denotes the negative bag of words in the i^{th} document;

$$d_1 = d_1^+ + d_1^-$$

$$d_2 = \{d_2 - d_1^-\} - d_2^- = d_2^+$$

$$d_3 = \{d_3 - d_1^- - d_2^-\} - d_3^- = d_3^+$$

$$..$$

$$...$$

$$....$$

$$d_n = \{d_n - d_1^- - d_2^- ... - d_{n-1}^-\} - d_n^- = d_n^+$$

Now, we can write the two Vocabularies as:

$$Vocab^+ = \Sigma d_i^+$$

$$Vocab^- = \Sigma d_i^-$$

Where $Vocab^+$ is the set of all postive words or the global positive words and $Vocab^-$ is the set of all negative words or the global negative words.

Final Vocabulary

After doing the above iteration for all the documents, we will have Vocabulary for all the documents. Finally we have two bags of words: global postives and global negatives. Figure below shows the interace where on the left side, we have all the postive words and on the right side we have all the negative words. We can transfer words from one bag to another by selecting them and click on the arrow button.

Chapter 5

Making Taxonomy and Tagging a Document

The making of a taxonomy or ontology is of great use. In this chapter, we will see, given a vocabulary and an input document, how to build an ontology. After that, we will reduce the set of MOOC-voc, i.e., online distance education vocabulary, into MOOC-tags by defining narrower and broader terms. Then, on the basis of frequency, we will give tags to the given input document.

5.1 Taxonomy or Ontology

Taxonomy/Ontology is the science of classification of things or concepts, including the principles that underlie such classification.
Taxonomy is a process of classifying content and organizing. It is an organized set of words used for organizing information which is intended for browsing. For faster information retrieval and better classification of knowledge, taxonomy is very much essential. The term Taxonomy comes from terms Taxos, ordering and nomos, rule. Taxonomy was first used in the field of biology where it was necessary for classification of biological specimens. Examples of taxonomies includes Blooms taxonomy, Plant taxonomy, and Animal taxonomy etc. which have been used today for easier classification of biological specimens. Nowadays the concept of taxonomy is being

used in other areas such as Psychology and Information Technology. Particularly in Information Technology, it is very much useful for content management and information architecture. This has been widely used in websites for categorization of web pages or resources (audio, video, content etc). Taxonomy is always rigid and conservative. Taxonomies also provide serendipitous guidance since it helps to get additional information from viewing where a topic resides in the taxonomys context. Many advantages are there in using taxonomy. Some of them include easy navigation and searching. However updating or maintaining taxonomy is very much difficult since incorporation of new resources or categories involves more time. There are three ways of constructing taxonomy: a manual approach, a semi-automated approach, an automated approach. From an organization perspective, taxonomy construction can be classified into three types namely buying pre-built taxonomy, building a taxonomy using several techniques and automatic approach. According to survey made by Gartner that taxonomy construction is vital and 70% of organizations who invested do not achieve their return on investment because of lack of proper taxonomy construction.

Importance of Taxonomy :

For any information to be organized, taxonomy is essential. Taxonomy plays a very important role for information and content management. Also it helps in searching of content. The most common method for constructing taxonomy was the manual construction. As the information available today is huge, constructing taxonomy for such information manually was time consuming and maintenance was difficult.

Constructuing Taxonomy :

Construction of taxonomy is limited to a particular domain. For example, taxonomy for a domain Sports can be constructed by specifying the categories Football, Cricket, Hockey etc under Sports. For extraction of categories and terms that can be used for each category, careful detailed analysis and study should be performed and this is defined by the domain experts. After a thorough analysis, the categories and content in each category are represented in an organizational structure. [2] As mentioned above taxonomies built using existing taxonomy templates (pre-built taxonomy) from vendors can speed up the construction of taxonomy and help an enterprise deliver quick results. Existing taxonomies can be optimized for the organizations specific requirements. However pre-built taxonomies have some disadvantages since it has less applicability and also time spent on user training.

An in-house constructed taxonomy is more particular to an organization and its intention. The selection of terminology in taxonomy is fully controlled by the developer. Sometimes it is only possible to construct an inhouse taxonomy since existing taxonomies may not exist for a particular domain. The only disadvantage for constructing taxonomy is time consumption and also expensive.

Irrespective of whatever approach used to construct taxonomy, there are four phases in general for taxonomy construction:

- Planning and Analysis: Detailed study needs to be done by the domain experts to identify the categories, resources to be allocated, cost involved in the construction.

- Design, Development and Testing: Detailed design of hierarchical structure is done by the software development team.

- Implementation: In this chapter, various approaches of implementing taxonomy are discussed.

- Maintenance: Maintenance of taxonomy is a taxing job and time consuming for manual construction as mentioned above. However maintenance can be simpler if automatic construction approaches is used.

For constructing taxonomy, two techniques are widely used: Top-Down approach and Bottom-Up approach.

- The top-down approach involves selection of few numbers of higher categories reaching more specific levels of lower subcategories based on the context. Usually taxonomy is developed manually and it provides control over the concepts present in higher taxonomy levels.

- The bottom-up approach involves selection of specific levels of categories and reaching the higher categories. To extract concepts from content and to make generalizations the automatic techniques are used in this approach.

The above two approaches have both advantages and disadvantages still vital for taxonomy construction.

5.2 Manual Approach

Usually the most common method of constructing taxonomy is the manual method. This method has been by the domain experts who are experienced in a particular domain can construct taxonomy. It provides major control over the synonyms and order of concepts. The choice of terminology is left to domain experts for using in taxonomy. Because of human judgment, manual classification of documents to the concepts in taxonomy is less accurate. Due to this misunderstanding of the terminology is possible for an end user who wants to view a particular resource of a domain. Also maintenance of taxonomy using such approach is a time consuming task. Nowadays it is very rare to construct taxonomy using manual approach. Advantages: Human decision, High precision, Disambiguation. Disadvantages: Labor exhaustive, Unable to scale, Costly resources.

5.3 Automatic Approach

In recent years research is being out for generating taxonomy using various techniques. Some of the approaches used for automatic Taxonomy generation include:

- Using WordNet (Lexical Database Dictionary) and NLP (Natural Language Processing) techniques.

- Using large text corpus.
- Clustering algorithms.
- Using the combination of tags (Annotations/Keywords) and Wikipedia to generate taxonomy.

The above approaches can be used in any combination for enhancing the construction of taxonomy. Also the above approaches are used at lexical and semantic level where the concepts of taxonomy are extracted and semantic relationships are used to construct the taxonomy.

Several automatic classification tools are available for classifying the content for a prevailing taxonomy or to generate taxonomy structure. Various algorithms (Statistical Analysis, Bayesian Probability, Clustering) [3] are applied to tools that create taxonomy structure to a set of documents using bottom-up strategy since this strategy involves incorporating automatic techniques. However automatic construction provides least control over the synonyms and order of concepts. Also refinement of the concepts is required for the user to understand. It can save time however human judgment will be there to check if the concept should be there in taxonomy or not. Advantages: Handles large volumes, Measures easily, Cheap resources Disadvantages: Rule/ algorithm weakness, Inaccuracies, Not easy to train

5.4 Constructing Taxonomy Using Tags

Social Tagging is a present trend now. People tag a resource which can be used for better sharing and searching. [11] Tagging helps in discovering items which are not found and helps in improving search. Several websites are available where people tag content or resources for effective communication. Some of them include Flickr, **Delicious, Bibsonomy, Technorati which are widely used portals for tagging. Basically tagging can be represented as Documents, Users and Tags. [14] This is sometimes known Collaborative Tagging. Since tags are used to describe the resources, to categorize the resources into a structured hierarchy tags play an important role for generating Taxonomy. This section describes the approaches used to create taxonomy from user generated tags. Also it gives an overview about the problems that can occur from user generated tags.

Some of the approaches include [12, 22] are used to construct Taxonomy. Reference [12] provides a framework to classify web pages based on social annotation. In this approach both web page and category are described based on tags and assign the resource to the category based on cosine similarity. Reference [13] describes a hierarchical classifier that can be used to classify documents into categories based on the tags that are used to describe the documents. This approach requires the document to be preprocessed before applying the document in the hierarchical classifier. Reference [14] describes about the document classification categorized using Open Directory9 . Reference [16] provides a novel approach for generating Taxonomy using tags. In this approach tags are collected from Delicious database and heuristic rule analysis is performed. Valid documents are extracted using tags with the help of Wikipedia. Each document is parsed and conceptrelationship acquisition and inference approach is performed for generating Taxonomy. Another approach proposed by [22] where tags are extracted from repositories and clustering techniques are performed. Similarity between tags can be calculated by using the distance metric which depends on the factors namely: Co-occurrence for tags and Semantic similarity for tags. Presently research is being carried out for enhancing taxonomy construction with the help of tags and also improving the navigation of Taxonomy. Tagging provides an easier approach for classification of content and constructing Taxonomy. However tags can be misused since it is user generated data. The vocabulary of tag terms may not be accurate. Also spams are generated using tags which are being addressed as a serious issue [11].

5.5 Our Approach

We have been using the Stanford Depedencies Parser API for making out the dependencies between different words of a sentence in a document.

The Stanford typed dependencies representation was designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used by people without linguistic expertise who want to extract textual relations. In particular, rather than the phrase structure representations that have long dominated in the computational linguistic community, it represents all sentence relationships uniformly as typed dependency relations. That is, as triples of a relation between pairs of words, such as the subject of distributes is Bell. Our experience is that this simple, uniform representation is quite accessible to non-linguists thinking about tasks involving information extraction from text and is effective in relation extraction applications.
Here is an example sentence:

Bell, based in Los Angeles, makes and distributes electronic, computer and building products.

For this sentence, the Stanford Dependencies (SD) representation is:
nsubj(makes-8, Bell-1)
nsubj(distributes-10, Bell-1)
vmod(Bell-1, based-3)
nn(Angeles-6, Los-5)
prep in(based-3, Angeles-6)
root(ROOT-0, makes-8)
conj and(makes-8, distributes-10)
amod(products-16, electronic-11)
conj and(electronic-11, computer-13)
amod(products-16, computer-13)
conj and(electronic-11, building-15)
amod(products-16, building-15)
dobj(makes-8, products-16)
dobj(distributes-10, products-16)
These dependencies map straightforwardly onto a directed graph representation, in

which words in the sentence are nodes in the graph and grammatical relations are edge labels. Figure 1 gives the graph representation for the example sentence above.

The current representation contains approximately 50 grammatical relations. The dependencies are all binary relations: a grammatical relation holds between a governor (also known as a regent or a head) and a dependent. The grammatical relations are defined below, in alphabetical order according to the dependencys abbreviated name (which appears in the parser output). The definitions make use of the Penn Treebank part-of-speech tags and phrasal labels.

5.6 Tagging with MOOC-tags

In the fourth chapter, we described how to build up the vocabulary from the given set of documents. This vocabulary consists of approximately 1000 keywords. Now the main task here is to reduce these keywords to some tags which would be less in number and then we will tag any given document by these tags. We will follow the following basic steps in order to tag a given document with the MOOC-tag.

- For a given document we will first convert it into the text file, tokenize, do stemming, remove stop words and finally we will have the set of all words in the document. The details of all these steps are given in chapter four of this thesis.
- After getting all the words in the document, we will take intersection of these words with the keywords in the vocabulary.
- Now, at this point of time we are having all the vocabulary keywords ocurring in the document with their corresponding frequency
- After this we will replace these keywords present in the document with their corresponding tag and thus update the frequency.
- Now, on the basis of these tags, we will tag a given input document.

Chapter 6

Conclusion and Future Work

In this chapter, we will summarize the work done in this thesis. We will also describe the Future work which can be done in this field.

6.1 Concluding Points

- We have generated a vocabulary from around 360 Online Distance Education Documents. Vocabulary consists of around 950 keywords rectified by the domain experts.

- In making of the taxonomy from the MOOC documents and given vocabulary we have used the Stanford Dependencies Parser API, working of which has been described vividly in the fifth chapter. Finally, it gives the hierarchical graph showing all the relationship between the different keywords.

- In tagging the document, we have tagged it statistically, i.e., on the basis of frequency of the most ocurring word. From the MOOC- vocab, we have reduced them to MOOC-tags and finally we have tagged a particular document with these tags.

6.2 Future Work

Keyword extation and tagging is a wide area in today's world and a fertile area in the field of Natural language Processing as well. There is a lot of research one can do in this field.

Following are the main works which one can add to extension of this thesis work:

1. To build more refined vocabulary which can take account of greater coverage of area. In this work, we are building vocabulary from around 360 Online Distance Education Documents. If we had taken more documents (say 1000 documents), then our vocabulary would have been more robust and complete.

2. While selecting the postive word from the set of all positive words, one can think of some automatic approach to do this work. At least we can reduce this set of possible words to less number of words to make it easy for the domain expert.

3. To make our Vocabulary more robust and reliable, we should show it to several domain experts and take feedback also. For making a good workable vocabulary we should do this iterations several times.

4. The Taxomony/Ontology which we are building for a given document can be more semantic in nature. Right now we are using Stanford Dependencies Parser API which is taking account of grammatical relationships and thus make the parse tree. To improve the Ontology relationships between the different keywords, more ideas can be think of in this direction.

5. The automatic tagger algorithm can be improved. In our work, we are doing the Statistical tagging which is based only on the number of terms ocurring (i.e., frequency) in the document. We can take into account the postion of word, bi-tri grams to improve the algorithm accuracy.

Bibliography

[1] http://www.internetworldstats.com/.

[2] Koraljka Golub. *Using Controlled Vocabularies in Automated Subject Classification of Textual Web pages, in the context of browsing.* TCDL Bulletin, Volume 2 Issue 2, 2006.

[3] Elaine G. Toms Shigeo Sugimoto Edie M. Rasmussen, Ray R. Larson, editor. *ACM/IEEE Joint Conference on Digital Libraries, JCDL,* Vancouver, BC, Canada, June 18-23 2007.

[4] Hend S. Al-Khalifa and Davis Hugh C. Folksonomies Versus Automatic Keyword Extraction: An Empirical Study. *International Journal on Computer Science and Information Systems Vol 1, No. 2, pp. 132-143, ISSN 1646-3692, 2006.*

[5] Hugh C. Davis Hend S Al-Khalifa. FasTa: A Folksonomy-Based Automatic Metadata Generator. *Lecture Notes in Computer Science,* 4753/2007, 2007.

[6] M Mahoui S Jhones. Hierarchical document clustering using automatically extracted keyphrases. In *Proc of 3rd International Asian Conference on Digital Libraries,* 1992.

[7] C.H. A. Koster A.T. Arampatzis, T. Tsoris and T.P. van der Weide. Phrase-based information retrieval. In *Information Processing and Management 34(6), 693707,* 1998.

[8] A Hulth. *Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction.* PhD thesis, Computer and Systems Sciences, Stockholm University, 2004.

[9] G. Paynter I. Witten C. Nevill-Manning Gutwin, C. and E. Frank. Improving browsing in digital libraries with keyphrase indexes. In *Technical report, Department of Computer Science, University of Saskatchewan, Canada,* 1998.

[10] Xiao Yu Jiang. Chinese Automatic Summarization Based on Keyword Extraction. *First International Workshop on Database Technology and Applications,*ISBN 978-0-7695-3604-0, April 25- April 26 2009.

[11] Black WJ Johnson FC, Paice CD. The Application of linguistic processing to automatic keyword generation. *Journal of Document and Text Management,* 1:215241, 1993.

[12] R. Olshen C. Stone L. Brieman, J. Friedman. Classification and Regression Trees. In *Wadsworth.* 1984.

[13] Brieman. Bagging Predictors. In *Machine Learning*, volume 24, pages pp 123140. 1996.

[14] P.D.Turney. Learning to extract keyphrases from text. Technical report, National Research Council, Institute for Informationn Technology, 1999.

[15] Ian H. Witten Carl Gutwin Craig G. Nevill-Manning Eibe Frank, Gordon W Paynter. Domain-Specific Keyphrase Extraction. In *Proceedings of the Sixteenth International Joint Conference on Artifical Intelligence*, pages 668673. Morgan Kaufmann Publishers Inc, San Francisco, USA, 1999.

[16] Martin Zaidel Daniel Karp, Yves Schabes and Dania Egedi. A Freely Available Wide Coverage Morphological Analyzer for English. In PROC of COLING, Aug 23-28 1992.

[17] Evelyne Tzoukermann Judith Klavans, Christian Jacquemin. A Natural Language Approach to Multi-word Term Conflation. Columbia University, IRIN, IUT de Nantes, Bell Laboratories, Lucent Technologies.

[18] Starting OpenOffice.org as a service. http://artofsolving.com/node/10, [Visited on 9 Jan 2010].

[19] http://www.slais.ubc.ca/courses/libr517/02-03-wt2/projects/dewey/P1Section1.htm.

[20] Robert D. Rodiguez. Kaisers Systematic Indexing. Library Resources and Technical Services, 1984.

[21] Hend S. Al-Khalifa and Davis Hugh C. Folksonomies Versus Automatic Keyword Extraction: An Empirical Study. In International Journal on Computer Science and Information Systems Vol 1, No. 2, pp. 132-143, ISSN 1646-3692, 2006.

[22] Cyril W. Cleverdon. Aslib Cranfield Research Project: report on the testing and analysis of an investigation into the comparative effeciency of indexing systems. Staff Publications-Cranfield Library, Oct, 1962. hdl.handle.net/1826/836.

[23] H.P. Luhn. Keyword in Context Index for Technical Literature(KWIC Index). Yorktown Heights, 1959. N.Y.: IBM.