a powerful way of summarizing the information relevant to a user's needs. In the case of a domain specific search engine, the automatic identification of important information can increase accuracy and efficiency of a directed search. There is use of hidden Markov models (HMMs) to extract the relevant fields of research papers. These extracted text segments are used to allow searched over specific fields and to provide effective presentation of search results and to match references to papers. For example, noticing the pop-up ads on any websites showing the recent items you might have looked on an online store with discounts. In Information Retrieval two types of models have been used (McCallum and Nigam, 1998) [77]. Both modules assume that a fixed vocabulary is present. But in first model a document is generated by first choosing a subset of vocabulary and then using the selected words any number of times, at least once without any order. This is called Multi-variate Bernoulli model. It takes the information of which words are used in a document irrespective of number of words and order. In second model, a document is generated by choosing a set of word occurrences and arranging them in any order. This model is called multi-nominal model, in addition to the Multi-variate Bernoulli model, it also captures information on how many times a word is used in a document.

Discovery of knowledge is becoming important areas of research over the recent years. Knowledge discovery research use a variety of techniques to extract useful information from source documents like *Parts of Speech (POS) tagging*, *Chunking or Shadow Parsing*, *Stop-words* (Keywords that are used and must be removed before processing documents), *Stemming* (Mapping words to some base for, it has two methods, dictionary-based stemming and Porter style stemming (Porter, 1980) [103]. Former one has higher accuracy but higher cost of implementation while latter has lower implementation cost and is usually insufficient for IR). *Compound or Statistical Phrases* (Compounds and statistical phrases index multi token units instead of single tokens.) *Word Sense Disambiguation* (Word sense disambiguation is the task of understanding the correct sense of a word in context. When used for information retrieval, terms are replaced by their senses in the document vector.)

The extracted information can be applied for a variety of purposes, for example to prepare a summary, to build databases, identify keywords, classifying text items according to some pre-defined categories etc. For example, CONSTRUE, it was developed for Reuters, that is used in classifying news stories (Hayes, 1992) [54]. It has been suggested that many IE systems can successfully extract terms from documents, acquiring relations between the terms is still a difficulty. PROMETHEE is a system that extracts lexico-syntactic patterns relative to a specific conceptual relation (Morin,1999) [89]. IE systems should work at many levels, from word recognition to discourse analysis at the level of the complete document. An application of the Blank Slate Language Processor (BSLP) (Bondale et al., 1999) [16] approach for the analysis of a real-life natural language corpus that consists of responses to open-ended questionnaires in the field of advertising.

There is a system called MITA (Metlife's Intelligent Text Analyzer) (Glasgow et al. (1998) [48]) that extracts information from life insurance applications. Ahonen et al. (1998) [1] suggested a mainstream framework for text mining that uses pragmatic and discourse level analyses of text.

e) Summarization

Overload of information is the real thing in this digital age, and already our reach and access to knowledge and information exceeds our capacity to understand it. This trend is not slowing