Using Bag of Words (BOW) and Standard Deviations to Represent Expected Structures for Document Retrieval: A Way of Thinking that Leads to Method Choices

By H.S. Hyman, JD, MBA, of University of South Florida, IS/DS Department, and
Warren Fridy III of Nova Southeastern University

Abstract/Summary

This paper discusses a theory and proposed design for a retrieval artifact using a bag of words (BOW) approach for terms and a standard deviation method for assigning weights. The paper is organized into three parts. The first part is an historical overview of the development of text mining techniques. It is intended to acquaint the reader with our perspectives and assumptions; it is not intended to be an exhaustive review of the literature. The second part discusses the application of text mining techniques to the legal domain, specifically eDiscovery. The third part describes our approach to the 2010 TREC Legal Track problem set #301.

Part Three is sub-divided into three sections. Section one is a discussion of our approach to document retrieval. Section two is a description of our design approach and method specifically chosen for Problem #301. Section three discusses contributions, limitations, generalizability, and conclusions based on our experience with the initial results produced by the competing artifacts in the TREC proceeding. We include a discussion of the initial results as reported at the conference.

Introduction

eDiscovery is an emerging problem domain that calls for solutions provided from two separate disciplines: Law and Information Systems (Hyman, 2010). The term eDiscovery refers to electronically stored information (ESI) sought by an opposing party during litigation (TREC, 2008 Proceedings).

For many years, lawyers and their clients have relied upon manual and physical methods for retrieving and providing requested documentation during the litigation process. These methods served their purpose when documentation was largely hard copy paper and stored almost entirely within physical files and storage containers.

As ESI began to have an increased share of the percentage of documentation sought in litigation, lawyers and their clients were forced to develop new approaches for seeking and providing such electronic documentation.

Early approaches to electronic discovery can be described as a “don’t ask, don’t tell” agreement between the parties —“I won’t ask it from you, if you don’t ask it from me.” In many, if not most instances, lawyers have been intimidated by the prospect of eDiscovery largely due to several challenges that must be considered: There is no bright-line legal precedent in this area, so how should I advise my client? What documents am I looking for? How do I ask for them? How can I force the other party to comply? What arguments can I make to the court to assist me in forcing the other party to comply—will the court be as intimidated as the lawyers when drafting an order?

Sometimes it is the clients who are intimidated by the prospect of eDiscovery. If there are 15 Gigs of information out there, and it costs a client $100 - $500 per Gig depending on the vendor, that can add up to a lot of money to reserve in a discovery budget already committed to depositions, process service, copies, and other ancillary items. As a result, many lawyers report that their clients are willing to “leave money on the table,” rather than litigate a commercial dispute, that they may or may not win an award, that may or may not get reduced by the court, that may or may not get reversed upon appeal.

We believe these issues remain relevant today. It is important to think about the problem set in these terms, especially given this year’s provision for acceptance of manual retrieval as a viable approach (TREC 2010 Legal Task Interactive Guidelines), and for one of the problem sets dedicated to privilege claims (Problem #304). This perspective is what drives our philosophy and our approach to the TREC 2010, Problem #301.

Historical Overview

There have been studies as early as the 1950s comparing automated methods for classification of documents (Hyman, 2010). Early methods were founded upon NLP, natural language processing (Hyman, 2010). Zellig Harris conducted studies on linguistics in the 1950s and published a paper entitled “Distributional Structure” in the journal *Word* in 1954. In 1964, Harold Borko published findings from an experiment comparing automated classification to human classification.

In the late 1960s, Goffman proposed an “indirect method” for document retrieval. The method is based on a theory of conditional relevance. The process takes a set of potentially relevant documents and assigns a probability to each document by evaluating the probability that document ‘X’ is relevant, if it is given that document ‘Y’ is relevant. (Croft and Van Rijsbergen, 1976).[1]

In the 1980s we see greater development of probabilistic approaches and theories based on semantics. In 1986 Marcia Bates published a model for searching unstructured documents. The model design incorporated three principals: Uncertainty, Variety, and Complexity.

In 1990, Deerwester, Dumais, Furnas, and Landauer published the seminal work in the search approach known as Latent Semantic Analytics (LSA). According to Huang and Kuo, LSA[2] has been a method proposed to deal with the issue of the poor fit associated with the limitation of synonyms, polymorphisms, and dictionaries.[3]

Ascertaining colloquial slang in an industry, sub-industry, or among operators in the field can be a difficult task. Many references in documents are based on local jargon often never considered by the drafter for the purpose of informing personnel outside the technical team loop. Hence the problem with relying upon industry dictionaries such as UMLS (Unified Medical Language System) in the medical domain, or the emerging LOIS project (Legal Information Sharing) in the legal domain.

[1] This method should sound familiar to TREC participants in the Interactive Task – taking suspect documents ‘Y’, presenting them to the topic authority, receiving feedback regarding the documents, and using that feedback to assess the relevance of documents ‘X.’

[2] Also referred to as Latent Semantic Indexing (LSI).

[3] TREC participants can certainly relate to the difficulties encountered when proposing search terms.

The nature of document search is iterative. What search operator would feel confident performing a single pass and declaring *that* single set of results to be reliable? Initial search terms need to be refined. Parameters need to be re-evaluated. This brings us to query expansion.

In 2008, Wang and Oard published an article experimenting with three classes of “vocabulary enrichment.” The classes were interactive relevance feedback, fully automatic relevance feedback, and thesaurus expansion. Their results supported their criticism of query expansion as “noisy.”

Also in 2008, Schweighofer and Geist proposed a method that made use of “lexical ontologies” and user feedback. They made use of a document oriented database *askSam* and an open text tool server Solr. Their implementation was based on a user generated query that is expanded based upon an ontology and relevance feedback. They formulated an iterative loop creating the new expanded query. Their approach is an excellent example of combining a first pass operator list of terms, augmented by a dictionary, and using feedback in an iterative loop. Their approach addresses a significant issue with unstructured document/text mining, combining industry terms (a dictionary) with colloquial slang and relevance feedback. This can be implemented with an iterative loop that addresses two challenges in unstructured document/text mining, query expansion and probabilistic relevance.

Text Mining in the Legal Domain

As of 2010, eDiscovery still represents a new and emerging challenge in the legal IR domain. The legal domain consists of complex terms and abstract concepts not easily represented (Scheighofer and Geist, 2008). eDiscovery is a complete departure from the traditional reasons for using legal IR.

Traditionally lawyers have been used to searching for documents in the form of cases using key notes, an NLP/BOW (bag of words) approach. This method was well suited to the goals of the legal domain – searching with static terms for cases that established legal precedent. The method of search relied upon a manual indexing -- cases announced in court reporters would be painstakingly catalogued by key note. This was done specifically so that documents catalogued could be retrieved at a later time. Compare this cataloguing to whatever method may be in place to catalogue a potential group of unknown number of documents sought in litigation; the challenge becomes apparent.

The documents retrieved using key words are cases that allow lawyers to be informed on a subject matter and predict a possible range of outcomes for a client's case. An example of this would be a lawyer who wants to advise a client who has recently been arrested for the possession of drugs resulting from a search of his car during a traffic stop. The lawyer could use one of the commercially available legal search engines and enter a simple key word search: automobile, drugs, seizure. The result produced would most likely be a reliable number of cases that the lawyer could use to advise the client about issues involved in the case and the range of outcomes.

By contrast, a search for discovery documents has no such reliable static terminology. The documents existing in whatever industry being litigated contain dynamic, not static terminologies. Quite often terms, definitions and references vary from company to company, and division to division within a company. It is no wonder why many practitioners may choose to simply avoid the problem completely through an agreement that, "I won't ask for your eDocuments, if you don't ask for mine."

Given the above we believe that one of the foremost concerns of eDiscovery is to determine methods to search for documents that were never intended for later, third party retrieval. Searching for a case using the key term *automobile* has a very strong likelihood of capturing cases involving trucks, cars, motorcycles and most other forms of personal conveyance. The same cannot be said of searching for all documents in a person's hard drive that are associated with the term "oil" or "petroleum."

There are two foundational issues that must be addressed when searching an unstructured domain. The first is the synonym problem. When a user searches for an automobile, the search engine needs to include car, truck, motorcycle, etc. The second problem is known as "polysemy," many words having more than one meaning. (Deerwater, et al. 1990). Synonyms and polysemies are two factors that reduce the power and accuracy of IR.

We observed in the TREC 2010 problem sets that a formidable challenge in eDiscovery is the nature of the documents themselves. Often, if not always, the documents sought during litigation have been created with no consideration for future, third party search. Many times documents are quick streams of thought jotted down and emailed to fellow workers. Quite often subject headings are misleading or attachments are mislabeled, both of which can result in a relevant document being missed or a non-relevant document being included.

Gap in Current Research and Our Approach

The current approaches to IR devote the majority of resources to complex classifier design. Our focus is on the primary purpose of IR – the successful retrieval of a maximum number of relevant documents to the user. To achieve this we advocate a "back to basics" approach to determine if recall success is being impacted by the complexity of the classifier or by the nature of the data set itself. We have framed our approach using hypotheses listed in the next section.

Hypotheses

H1: BOW terms manipulated through iteration and weighting methods will produce recall results as good as a sophisticated classifier.

H2: The use of mean occurrences and standard deviations will produce precision results as good as a sophisticated classifier.

Our Design Artifact and Method for Problem #301

Our approach to design of a retrieval system focuses on choice points along the search query building process. The first choice is the collection of terms to be used. The second choice is the weighting method and assignment of weights to the terms and occurrences. Choices have direct consequences to the width and depth of the net cast. These consequences should be taken seriously by the user and the system designer alike.

We suggest the assignment of weights should be based on the structure of the terms one would expect to see in a target document. In theory, the collection of target documents should closely resemble the expected structure of terms. From the expected structure we predict a mean number of term appearances. In this case, we assign a threshold for acceptance based on the standard deviation of the expected terms from the mean appearances. There are numerous alternative threshold assignments that may be used. We believe the key factor is how the expected structure of selected terms appears in targeted documents.

Our experimental design in this case is a bag of words (BOW) approach, where the terms are weighted using standard deviations from the mean number of occurrences of terms in the suspect document. Our first step is the development of a dictionary of terms that on their face are good candidates for inclusion, given the subject area and domain. The next step is to gather and review synonyms and colloquial slang. The initial terms, synonyms and slang, form the first pass query. The first pass is implemented with no weights. The results produced from the first pass are analyzed by mean number of appearances per term and standard deviation. We find that user choices for the construction of the terms and the mechanism for scoring the relative importance of a particular term's appearance, impact the final set of documents targeted.

The nature of the domain searched and the use of language in that domain should not be underestimated. When a set of terms is agreed upon, some form of article construction along with an assigned weight is applied. For example, we have found through simple trial and error that a given term appearing several times in a document may be a good sign, and should receive a relative score to reflect such. However, a term that appears too many times could indicate an irrelevant document, and should receive a corresponding relative score to reflect such. We assign weights to the number of appearances of terms in a document. The choice of weighting has a direct impact on whether a document is included or excluded.

Our design assumes the purpose behind the first pass to be to capture as many documents as possible. The main concern at this stage is missing a document due to failure of synonyms.

The purpose behind the second pass in our design is to narrow the documents returned. It is for this reason that we use a more targeted group of terms, comprised mostly from the collection of initial terms.

We implement an evaluation stage to validate the structural pattern of terms. There are variations that can be implemented at this choice point. One variation is to return the group of documents based on a count of all terms occurring in a document. Another variation is to return documents based on per term count and then weight each term according to its count. A third variation is to count all occurrences of nouns in the document. This is done similar to a grammar checker. A fourth variation is to return a word list. This list would contain all newly discovery terms in a document. What makes this option appealing is the ability to have the user review the pattern of terms for possible document elimination. In other words, if a suspect document has a particular pattern of terms, eliminate it. This method allows the system to account for the misuse of terms, and the overuse of terms.

Calculations based on counts of terms in the documents must be assigned whichever method is chosen for the evaluation stage. These calculations lead to scores for documents to be grouped for comparison. In this case we chose to use standard deviations from the mean number of appearances for weights. Just as in evaluation choices, there are numerous alternative methods for calculating term occurrences and scoring documents. One such example as mentioned earlier is to count all terms, or count per term. Whichever method is selected, the nature of our approach lies in the choice points the user makes in the article construction and scoring mechanism. Our design and choice points based on our theory has been graphically depicted in Figure 1. For the problem #301 experiment we chose to use the first pass and second pass as described, and implement Variation 1 and Evaluation method 1 for third pass and output.

Data Set

The data set used in this case is a portion form the ENRON data set. It contains over 650,000 emails, some with attachments and some without. A document (email) can be deemed relevant due to the body of the email containing relevant information, or the attachment to the email containing relevant information. The data set has been previously validated by prior Text Retrieval (TREC) proceedings and is publically available through the Electronic Discovery Relationship Model (EDRM) web site.

Discussion of Results

The initial results suggest that our method produced the highest recall of documents (.20). However, our method also produced the lowest level of precision (.20). Our F-score result, which is a function of recall and precision, was the second highest (.20), despite our extremely low performance in precision.

How should these results be interpreted? Well, like many experimental methods, the results are circumstance driven. The data set comprised approximately 650,000 documents, of which we retrieved approximately 23,000. That represents a lot of documents for a user to review, especially with precision as low as 20%. Consider the fact that the next highest amount of documents retrieved was around 13,000, albeit with a higher precision rate. Which method has proven to be better: The method that produced more documents, but also more irrelevant documents, or the method that produced fewer documents, but a higher percentage of relevant ones? Success and failure when described this way becomes a question of customer satisfaction – elusive to define, and difficult to satisfy. We suggest that the utility of our experiment lies in its ability to compete effectively with the more sophisticated approaches applied to problem #301. The utility of our artifact lies in the ability of the user to make iterative improvements to the classifier based on human in the loop feedback. This has the effect of allowing the decision maker to continue evaluating test run documents, and use that feedback to improve the next iteration by the user.

Contribution

This paper contributes to the domain of unstructured document retrieval by offering a design artifact that allows a user to conduct, on the fly, retrieval experiments. Our artifact offers users the ability to choose terms, assign weights, and create a scoring mechanism. The user observes results based on choices made. The output of our system is a report in the form of an Excel spreadsheet in CSV format. The report affords the user the ability to select and filter, based upon columnar output, segregated by specific pass number. This output method allows the user the ability to add a human in the loop component to aid in recognizing emerging patterns that may escape an automated process. The power of our artifact lies in its simplicity. Our artifact can be used to benchmark results produced by more sophisticated classifiers that operate as black boxes.

Limitations

One limitation of this paper is the fact that the artifact designed is still largely a work in progress. Additional experiments need to be performed to better assess the impact of article construction, use of threshold values for term counts, and alternative choices for weighting methods.

A second limitation in this study focuses on the nature of retrieval terms themselves. The foundation of a search is based on the initial choices of terms and synonyms. If the initial terms are not well represented, even the best weighting system is not going to produce effective results.

Generalizability

Although the artifact in this case has limitations, the choices associated with the methods explored apply to all document retrieval problem sets and the decision to choose a particular classifier approach. We have found in our research that both the user and the designer of a retrieval artifact must give careful thought and consideration to the choice points discussed in this paper. Terms chosen, weights assigned, and word structure expected, all have impacts on the final result produced. Our artifact and experiment show one such example that we believe is prototypical in this domain.

Future Research

Additional experiments exploring the impacts of choice points associated with terms and weights will provide further insight for document retrieval in the unstructured domain. Further experimentation will also provide greater guidance regarding the utility of using a BOW approach as an alternative to classifiers such as Latent Semantic Indexing when choosing terms and weights for designing an artifact for document retrieval.

Conclusion

In this study we designed an artifact for document retrieval based on initial terms, synonyms and slangs. We chose standard deviation from mean number of appearances of terms in a document as a weighting method. We used a multiple step, iterative process to produce our results. We found choices made at several steps of the process have varied impacts on results produced. We found initial choice of terms and the expectation of how those terms are constructed in a document of interest had an impact on choice of weights applied and results produced.

The initial results support our hypothesis that a BOW approach can produce as good a recall as the more sophisticated classifiers. The initial results do not support our hypothesis that a BOW approach can produce as good a precision as the more sophisticated classifiers. Further experimentation is recommended to explore the significance of the impact of the choices made in designing the retrieval artifact, and determining how recall and precision are affected by the choice of classifier used.

References

M. Bates, “Subject Access in Online Catalogs: A Design Model,” *Journal of the American Society for Information Science*. (Nov. 1986)

H. Borko, “Measuring the Reliability of Subject Classification by Men and Machines,” *American Documentation*. (Oct. 1964).

W.B. Croft and C. J. Van Rijsbergen, “An Evaluation of Goffman’s Indirect Retrieval Method,” *Information Processing and Management*. Vol. 12, Pg. 327 (1976).

S. Deerwester, S. T. Dumais, G.W. Furnas, T. K. Landauer, R. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*. (Sep. 1990).

Z. Harris, "Distributional Structure". *Word*. Vol. 10, Pg. 146–62 (1954).

Chua-Che Huang and Chia-Ming Kuo, “The Transformation and Search of Semi-Structured Knowledge in Organizations,” *Journal of Knowledge Management*. Vol. 7, Iss. 4 (2003).

H.S. Hyman, “Designing a Text Mining Artifact for eDiscovery Solutions.” Working Paper (Submitted June 2010).

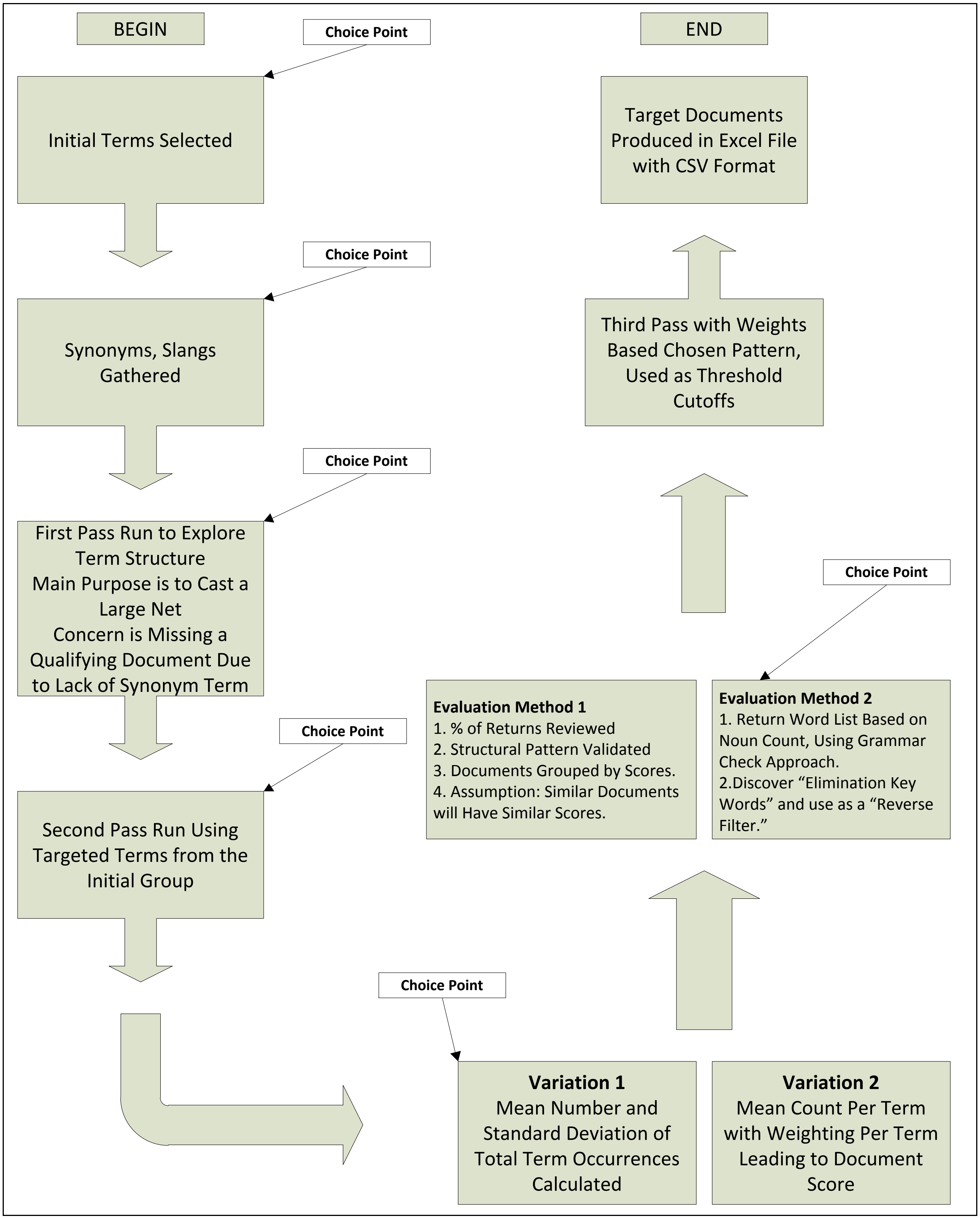
D. Oard, B. Hedin, S. Tomlinson, J. R. Baron, “Overview of the TREC 2008 Legal Track.” TREC Conference 2008, Proceedings.

E. Schweighofer and A. Geist, “Legal Query Expansion Using Ontologies and Relevance Feedback,” TREC Conference 2008, Proceedings.

C. J. Van Rijsbergen, D. J. Harper, M.F. Porter, “The Selection of Good Search Terms,” *Information Processing and Management*. Vol. 17, Pg. 77-91 (1981).

L. Wang, D. Oard, “Query Expansion for Noisy Legal Documents,” TREC Conference 2008, Proceedings.

Figure 1


BEGIN
Choice Point
Initial Terms Selected
Choice Point
Synonyms, Slangs Gathered
Choice Point
First Pass Run to Explore Term Structure
Main Purpose is to Cast a Large Net
Concern is Missing a Qualifying Document Due to Lack of Synonym Term
Choice Point
Second Pass Run Using Targeted Terms from the Initial Group
Choice Point
Variation 1
Mean Number and Standard Deviation of Total Term Occurrences Calculated
Variation 2
Mean Count Per Term with Weighting Per Term Leading to Document Score
Evaluation Method 1
1. % of Returns Reviewed
2. Structural Pattern Validated
3. Documents Grouped by Scores.
4. Assumption: Similar Documents will Have Similar Scores.
Choice Point
Evaluation Method 2
1. Return Word List Based on Noun Count, Using Grammar Check Approach.
2.Discover “Elimination Key Words” and use as a “Reverse Filter.”
Third Pass with Weights Based Chosen Pattern, Used as Threshold Cutoffs
Target Documents Produced in Excel File with CSV Format
END