DATA MINING INFORMATION SEARCH

Mr. Prashant S. Gawande[1], Prof. Vivekkumar Shrivastav[2]

[1]Information Technology Department,
Institute of Technology and Management,
Bhilware, Rajsthan-311001

[2]Assistant Professor,
Head of the Information Technology Department,
Institute of Technology and Management,
Bhilware, Rajsthan-311001

Abstract: Data mining is a process of collecting and analyzing the data for different purpose. Now a days Data mining is not only used in industries but it is also required for educational institutes. Data mining is an effective tool for decision making, cost cutting, increasing revenues etc. Information can be changed into knowledge and it can become a good source for any organization to survive in today's cut-throat competition. Most users take the help of search engines and browsers for obtaining data. However, the data we get from these sources is not ready to use type of data. It is very herculean task to covert these data into accurate information. It is just like searching diamonds in the huge ocean. This paper tries to give a new look to traditional data mining process. Web mining embodies three parts i.e. web structure mining, web content mining, web usage mining. This paper suggests a new framework or text mining based on the integration of Information Extraction (IE). Traditional data mining assumes that the information to be "mined" is already in the form of a relational database. Web mining deals with three main areas: web content mining, web usage mining and web structure mining.

Keywords: Text Mining, Extraction, Classification, Stemming, Stopword Removal

1. INTRODUCTION

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. Data mining consists of three basic steps Extract the information, load the information & display the information (out-put).

Text mining is the technique that helps users to find useful information from a large amount of digital text documents on the Web or databases. It is therefore crucial that a good text mining model should retrieve the information that meets user's needs within a relatively efficient time frame.

A first step toward any Web-based text mining effort would be to collect a significant number of Web mentions of a subject. Thus, the challenge becomes not only to find all the subject occurrences, but also to filter out just those that have the desired meaning.

2. TEXT MINING AND INFORMATION EXTRACTION

"Text mining" is used to describe the application of data mining techniques to automated discovery of useful or interesting knowledge from unstructured data.

Unstructured data exists in two main categories: bitmap objects and textual objects. Bitmap objects are non-language based (e.g. image, audio, or video files) whereas textual objects are "based on written or printed language" and predominantly include text documents.. Text mining is the discovery of previously unknown information or concepts from text files by automatically extracting information from several written resources using computer software [An evaluation of unstructured Text Mining software].

Text mining on Web adoptive technique include classification, clustering, associate rule and sequence analysis etc.. Among them, classification is a kind of data analysis form, which can be used to gather and describe important data setIn Web text mining, the text extraction and the characteristic express of its extraction contents are the foundation of mining work, the text classification is the most important and basic mining method.

2.1 Extraction

In extraction process, required information is extracted by checking maximum text density from the text contents from a web page. By this process, noise from the web page is removed. Extraction is followed by pre-processing of the text content. Pre-processing of the text contents include stemming and stop word removal.

2.2 Classification

In classification process, required information classified on the basis of classifier to which belong to documents.

3. GLOBAL INFORMATION

Unlike the Term Count Model, Salton's Vector Space Model [1] incorporates local and global information.

$$w_i = tf_i * log\left(\frac{D}{df_i}\right) \text{ ………..………. (1)}$$

Equation 1: Term Weight =

where

- tf_i = term frequency (term counts) or number of times a term i occurs in a document. This accounts for local information.
- df_i = document frequency or number of documents containing term i
- *D = number of documents in a database.*

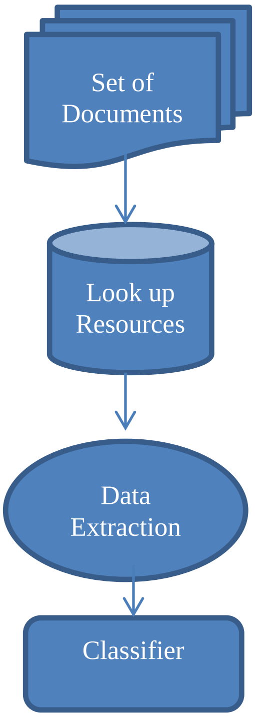


Figure 1 Web text mining Classification & Extraction

The df_i */D* ratio is the probability of selecting a document containing a queried term from a collection of documents. This can be viewed as a global probability over the entire collection. Thus, the *log(D/df$_i$)* term is the *inverse document frequency*, *IDF$_i$* and accounts for global information. The following figure illustrates the relationship between local and global frequencies in an ideal database collection consisting of five documents D1, D2, D3, D4, and D5. Only three documents contain the term "CAR". Querying the system for this term gives an IDF value of log(5/3) = 0.2218.

Those of us specialized in applied fractal geometry recognize the self-similar nature of this figure up to some scales. Note that collections consist of documents, documents consist of passages and passages consist of sentences. Thus, for a term i in a document j we can talk in terms of *collection frequencies (Cf), term frequencies (tf), passage frequencies (Pf) and sentence frequencies (Sf).*

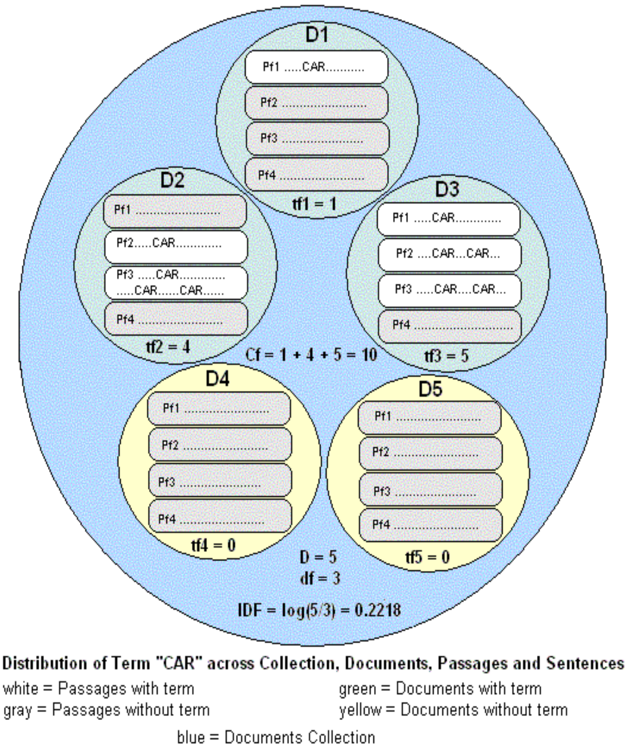

D1
Pf1 .....CAR...........
Pf2 ....
Pf3 ....
Pf4 ....
tf1 = 1
D2
Pf1 ....
Pf2.....CAR............
Pf3 .....CAR............ .....CAR......CAR......
Pf4 ....
tf2 = 4
D3
Pf1 .....CAR............
Pf2 ....CAR....CAR...
Pf3 .....CAR....CAR...
Pf4 ....
tf3 = 5
Cf = 1 + 4 + 5 = 10
D4
Pf1 ....
Pf2 ....
Pf3 ....
Pf4 ....
tf4 = 0
D5
Pf1 ....
Pf2 ....
Pf3 ....
Pf4 ....
tf5 = 0
D = 5
df = 3
IDF = log(5/3) = 0.2218
Distribution of Term "CAR" across Collection, Documents, Passages and Sentences
white = Passages with term
gray = Passages without term
green = Documents with term
yellow = Documents without term
blue = Documents Collection

Figure 1 Distribution Model –car

$$Cf_i = \sum_j tf_{i,j}$$

$$tf_{i,j} = \sum_p Pf_{i,j,p}$$

$$Pf_{i,j,p} = \sum_s Sf_{i,j,p,s} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Equation 2(a, b, c):

Equation 2(b) is implicit in Equation 1. Models that attempt to associate term weights with frequency values must take into consideration the scaling nature of relevancy. Certainly, the so-called "keyword density" ratio promoted by many search engine optimizers (SEOs) is not in this category.

3.1 Vector Space Example

To understand Equation 1, let use a trivial example. To simplify, let assume we deal with a basic term vector model in which we

1. do not take into account WHERE the terms occur in documents.
2. use all terms, including very common terms and stopwords.
3. do not reduce terms to root terms (stemming).
4. use raw frequencies for terms and queries (unnormalized data).

I'm presenting the following example, courtesy of Professors David Grossman and OphirFrieder, from the Illinois Institute of Technology [2]. This is one of the best examples on term vector calculations available online.

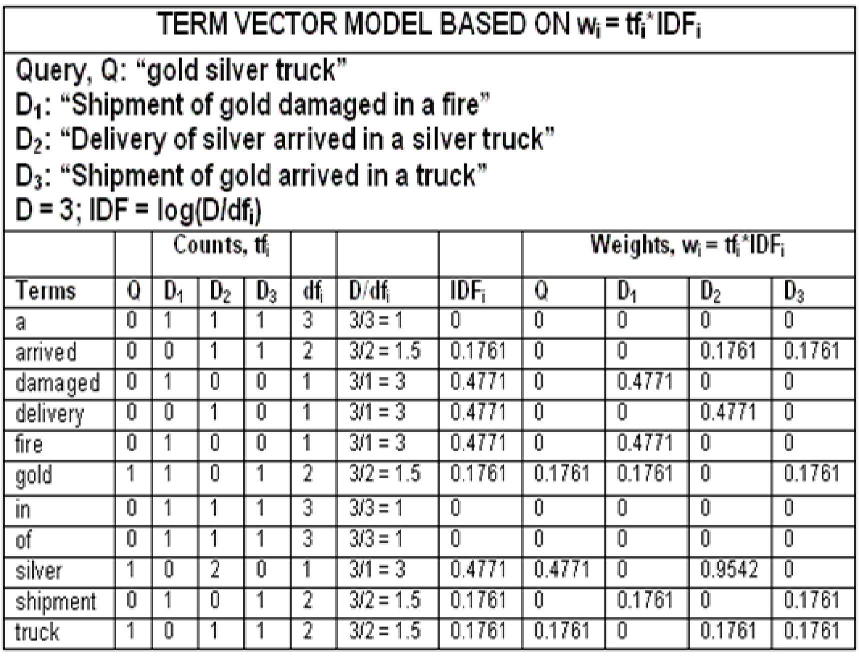
- By the way, Dr. Grossman and Dr. Frieder are the authors of the authority book *Information Retrieval: Algorithms and Heuristics*. Originally published in 1997, a new edition is available now through Amazon.com [3]. This is a must-read literature for graduate students, search engineers and search engine marketers. The book focuses on the real thing behind IR systems and search algorithms.

Suppose we query an IR system for the query "gold silver truck". The database collection consists of three documents (D = 3) with the following content

D1: "Shipment of gold damaged in a fire"
D2: "Delivery of silver arrived in a silver truck"
D3: "Shipment of gold arrived in a truck"

Retrieval results are summarized in the following Table 1.

Table 1 Retrieval Results



TERM VECTOR MODEL BASED ON $w_i = tf_i*IDF_i$

Query, Q: "gold silver truck"
D_1: "Shipment of gold damaged in a fire"
D_2: "Delivery of silver arrived in a silver truck"
D_3: "Shipment of gold arrived in a truck"
D = 3; IDF = log(D/df$_i$)

		Counts, tf_i						Weights, $w_i = tf_i*IDF_i$			
Terms	Q	D_1	D_2	D_3	df_i	D/df_i	IDF_i	Q	D_1	D_2	D_3
a	0	1	1	1	3	3/3 = 1	0	0	0	0	0
arrived	0	0	1	1	2	3/2 = 1.5	0.1761	0	0	0.1761	0.1761
damaged	0	1	0	0	1	3/1 = 3	0.4771	0	0.4771	0	0
delivery	0	0	1	0	1	3/1 = 3	0.4771	0	0	0.4771	0
fire	0	1	0	0	1	3/1 = 3	0.4771	0	0.4771	0	0
gold	1	1	0	1	2	3/2 = 1.5	0.1761	0.1761	0.1761	0	0.1761
in	0	1	1	1	3	3/3 = 1	0	0	0	0	0
of	0	1	1	1	3	3/3 = 1	0	0	0	0	0
silver	1	0	2	0	1	3/1 = 3	0.4771	0.4771	0	0.9542	0
shipment	0	1	0	1	2	3/2 = 1.5	0.1761	0	0.1761	0	0.1761
truck	1	0	1	1	2	3/2 = 1.5	0.1761	0.1761	0	0.1761	0.1761

The tabular data is based on Dr. Grossman's example. I have added the last four columns to illustrate all term weight calculations. Let's analyze the raw data, column by column.

1. Columns 1 - 5: First, we construct an index of terms from the documents and determine the term counts tf_i for the query and each document D_j.
2. Columns 6 - 8: Second, we compute the document frequency d_i for each document. Since $IDF_i = \log(D/df_i)$ and D = 3, this calculation is straightforward.
3. Columns 9 - 12: Third, we take the tf*IDF products and compute the term weights. These columns can be viewed as a sparse matrix in which most entries are zero.

Now we treat weights as coordinates in the vector space, effectively representing documents and the query as vectors. To find out which document vector is closer to the query vector, we resource to the similarity analysis introduced in Part 2.

3.2 Similarity Analysis

First for each document and query, we compute all vector lengths (zero terms ignored).

$$|D_1| = \sqrt{0.4771^2 + 0.4771^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.5173} = 0.7192$$
$$|D_2| = \sqrt{0.1761^2 + 0.4771^2 + 0.9542^2 + 0.1761^2} = \sqrt{1.2001} = 1.0955$$
$$|D_3| = \sqrt{0.1761^2 + 0.1761^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.1240} = 0.3522$$

$$\therefore |D_i| = \sqrt{\sum_i w_{i,j}^2}$$

$$|Q| = \sqrt{0.1761^2 + 0.4771^2 + 0.1761^2} = \sqrt{0.2896} = 0.5382$$

$$\therefore |Q| = \sqrt{\sum_i w_{Q,j}^2}$$

$$\text{Cosine } \theta_{D_1} = \frac{Q \bullet D_1}{|Q| * |D_1|} = \frac{0.0310}{0.5382 * 0.7192} = 0.0801$$

$$\text{Cosine } \theta_{D_2} = \frac{Q \bullet D_2}{|Q| * |D_2|} = \frac{0.4862}{0.5382 * 1.0955} = 0.8246$$

$$\text{Cosine } \theta_{D_3} = \frac{Q \bullet D_3}{|Q| * |D_3|} = \frac{0.0620}{0.5382 * 0.3522} = 0.3271$$

$$\therefore \text{Cosine } \theta_{D_i} = \text{Sim}(Q, D_i)$$

$$\therefore \text{Sim}(Q, D_i) = \frac{\sum_i w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \sqrt{\sum_i w_{i,j}^2}}$$

Finally we sort and rank the documents in descending order according to the similarity values

Rank 1: Doc 2 = 0.8246
Rank 2: Doc 3 = 0.3271
Rank 3: Doc 1 = 0.0801

3.3 Observations

This example illustrates several facts. First, that very frequent terms such as "a", "in", and "of" tend to receive a low weight -a value of zero in this case. Thus, the model correctly predicts that very common terms, occurring in many documents in a collection are not good discriminators of relevancy. Note that this reasoning is based on global information; ie., the IDF term. Precisely, this is why this model is better than the term count model discussed in Part 2. Third, that instead of calculating individual vector lengths and dot products we can save computational time by applying directly the similarity function.

$$\text{Sim}(Q, D_i) = \frac{\sum_i w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \sqrt{\sum_i w_{i,j}^2}}$$

4. CONCLUSION

By using computer, statistics, artificial intelligence we can try to present data in systematic way and that can be used for various purposes. In simple words data mining means discovering the undiscovered for effective decision making. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. Data mining consists of three basic steps Extract the information, load the information & display the information (out-put).

References

[1] Shiqun Yin Gang Wang YuhuiQiuWeiqunZhang.Research and Implement of Classification Algorithm on Web TextMining.IEEE2007.

[2] Shiqun Yin Yuhui Qiu1,ChengwenZhong . Web InformationExtraction and Classification Method .IEEE 2007

[3] Shiqun Yin YuhuiQiuJikeGe, XiaohongLan.Research andRealization of Extraction Algorithm on Web Text Mining.IEEE2007.

[4] JaideepSrivastava, PrasannaDesikan, Vipin Kumar "Web Mining— Concepts, Applications, and Research ".

[5] Ms. Sarika Y. Pabalkar"Web Text Mining for news by Classification" IJARCCE Vol. 1, Issue 6, August 2012