Quantification of Portrayal Concepts using tf-idf Weighting

S. Florence Vijila [1] and Dr. K. Nirmala [2]

[1] Research Scholar, Manonmaniam Sundaranar University, Tamil Nadu, India, Assistant Professor , CSI Ewart Women's Christian College, Melrosapuram , TamilNadu
[2] Associate Prof. of Computer Applications, Quaid-e-Millath Govt. College for Women, Chennai, Tamil Nadu, India

ABSTRACT

Term frequencies and inverse document frequencies have been successfully applied in determining weighting for document rankings. However these have been more successful in text mining and in extraction techniques used in the web. Concept mining has become increasingly popular in the research and application areas of Computer Science. This paper attempts to demonstrate the limited usage of term frequency and inverse document frequency for the application of weighting calculations for ranking documents that are based on concept quantifications. The case study considered for experiment in this paper, is based on concept terms of David Merrill's First Principles of Instruction (FPI). Merrill's FPI applies cognitive structures explicitly for analyzing instructional materials. Therefore it is justified that the terms categorized under each cognitive structure (or portrayal) of FPI can be taken as respective concept of that portrayal. As question papers are representative of cognitive structures in a more clear and logical way, four question papers on 'C Language' have been considered for the experimental study, that are detailed in this paper. Manual method has been adopted for the computation of quantities of portrayals in selected documents for the purpose of comparative study. As manual method is accurate, the values (results) are considered as benchmark values. These benchmark values are considered for comparing with normalized term frequencies that are derived (experimented) from automated extractions from the same selected documents. The study is however limited to four documents only. Conclusions are drawn from this experimental study, which will be of immense use to concept mining researchers as well as for instructional designers.

KEYWORDS

concept keywords; document ranking; term frequency weighting; cognitive structures.

1. INTRODUCTION

Literatures are aplenty for document ranking using term weighting. Documents are retrieved using keywords and documents are ranked according to keyword density (Masaru Ohba et al - 2005, SA-Kwang Song et al -2012). According to documented literature, ranking of documents based on term weighting has been a key research issue in information retrieval, particularly retrieving concepts in addition to keywords. Applying weighting schemes using term frequency and inverse document frequency are crucial for ranking of documents (Sa-Kwang Song et al-2012). A concept keyword is a word that represents a concept. But the nature of concept and the

meaning a particular keyword carries are subjective judgment, as per these authors. The authors have adapted term frequency for mining concepts. Both human selected concepts and automated methodologies have been successfully adopted. But research publications point out that concept mining ultimately depends only upon manual methods, when accuracy is needed to a high level. In other words, automated methods would only provide approximate results.

The tf-idf weighting (term frequency-inverse document frequency, a.k.a.TF-IDF) is a numerical statistic which reflects how important a word is to a document in a collection or corpus (Salton et al - 1988). By convention, the tf-idf value increases proportionally to the number of times a word appears in a document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. In view of this, the factor itself may be considered as a weighting factor in information retrieval and text mining. This paper hence attempts to determine the acceptability of tf-idf, through experimentation, on quantifying concepts from documents. Comparative study will be carried out from the results of manual methods.

In concept mining techniques, variations of the tf-idf weighting scheme are often used by some search engines as a central tool in scoring and ranking a document's relevance given by a user query. Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification. This paper attempts to quantify well proven conceptual terms, such as cognitive structures from instructional documents.

David Merrill (2007), in his 'First Principles of Instruction' (FPI) divides any instructional event into four phases, which he calls 'Activation', 'Demonstration', 'Application' and 'Integration'. Central to this instructional model is a real-time problem-solving theme, called 'problem'. Merrill suggests that fundamental principles of instructional design should be relied on and these apply regardless of any instructional design model used. Violating this would produce a decrement in learning and performance. Each phase is a cognitive portrayal for learning a concept according to this author. Therefore it would be very apt to consider cognitive structure or portrayal terms as concepts for the purpose of quantification.

We propose to rank order four sets of concept keywords of these four cognitive structures (portrayals) of Merrill's FPI. The rank ordering would be carried out using normalized term frequency (tf) and inverse document frequency (idf) on selected University question paper documents of "C Language'. To ascertain the degree of accuracy (the objective of this paper), we propose to compare the results with manual methods. Conclusions have been drawn from this experimental study. The results will be of immense use to concept mining methodologists and for instructional designers. The material of this paper forms a part of a whole research programme on concept based quantification of portrayals by the authors. As these portrayals are taken for concept quantification, the meaning of each portrayal is essential to comprehend before attempting to experiment with.

2. 'ACTIVATION' PORTRAYAL ON EVALUATION DOCUMENT OF 'C' LANGUAGE

The 'Activation' portrayal of FPI specifies, " Does the question direct students to recall, remember, repeat or recognize basic and fundamental knowledge of 'C' from the relevant past

behavior (experience) of the students that can be used as a foundation for the acquisition of new (and further advanced) knowledge of 'C'?".

Based on the above definition and on the past ten year question paper patterns analyzed by the authors, the following concept keywords have been specifically arrived at for the computation of term frequency:

"list, define, tell, name, locate, identify, what, give, distinguish, acquire, underline, relate, state, recall, select, repeat, recognize, reproduce and measure".

3. 'DEMONSTRATION' PORTRAYAL ON EVALUATION DOCUMENT OF 'C' LANGUAGE

The 'Demonstration' portrayal of FPI specifies, "Does the question directs the students to demonstrate of what has he learnt rather than merely provide information (unlike 'Activation') about what is needed as foundation for further techniques of 'C'?". Based on the above definition and the past ten year question papers pattern analyzed by the authors, the following concept keywords have been exclusively arrived at for the computation of term frequency:

> "demonstrate, summarize, illustrate, interpret, contrast, predict, associate, distinguish, identify, show, label, collect, experiment, recite, classify, discuss, select, compare, translate, prepare, change, rephrase, differentiate, draw, explain, estimate, fill in, choose, operate, perform, organize and write".

4. 'APPLICATION' PORTRAYAL ON EVALUATION DOCUMENT OF 'C' LANGUAGE

The 'Application' portrayal of FPI specifies, "Does the question provides an opportunity for student to apply her acquired knowledge of 'C' to solve a problem?". Based on this definition and from analysis carried out authors on the past ten year question papers, the following concept keywords have been exclusively arrived at for the computation of term frequencies:

> "apply, calculate, find, solve, illustrate, make, predict, construct, assess, practice, restrictive, classify, code, develop, generate, write".

5. 'INTEGRATION' PORTRAYAL ON EVALUATION DOCUMENT OF 'C' LANGUAGE

The 'Integration' portrayal of FPI specifies, "Does the question triggers and encourages the student to integrate (transfer) her acquired knowledge of 'C' to a new situation in any critical and complex situation?". Based on this definition and from the study on past ten year question papers, the following concept keywords have been arrived at exclusively for the computation of term frequencies:

> "analyze, solve, justify, infer, combine, integrate, plan, generalize, assess, decide, rank, grade, recommend, contrast, survey, examine, investigate, compose, invent, improve, imagine, hypothesize, predict, evaluate, rate, how and why".

Using these keywords, quantified benchmark values on the selected documents have been arrived at through manual methods. These benchmark values would be compared with automated term frequencies so as to arrive at conclusions.

6. COMPUTATION OF BENCHMARK VALUES BY MANUAL METHOD

Based on the definitions of the portrayals of FPI (four cognitive structures), quantification of concept terms in terms of these four cognitive structures on the selected four question paper documents (samples) were carried out using concept keywords as provided above. Thus, the addition of all values (in percentages) would yield to 100%, as the method adapted is pure manual by the authors. Therefore the values need to be exact and hence taken as benchmark values for comparative studies. The values are provided in Table 1.

Portrayal	Documents				Average
	Q1	Q2	Q3	Q4	
Activation	48%	27%	30%	31%	34%
Demonstration	22%	28%	42%	33%	31%
Application	26%	36%	21%	29%	28%
Integration	4%	9%	7%	7%	7%

Table 1. Benchmark values in percentages computed manually

7. RESULTS AND DISCUSSIONS ON WEIGHTING USING TERM FREQUENCIES

The four selected question papers (documents) of the subject 'C Language' have been used for the intended analysis. The normalized term frequency and inverse document frequencies on these selected four concept keywords, namely 'Activation', 'Demonstration', 'Application' and 'Integration' using their respective keywords have been used for the calculation of two frequencies. These term frequencies are used for determining the weighting of these four portrayals that we have computed through inverse document frequencies. The total number of words, the most frequented word and the number of concepts (different keywords of each concept or portrayal) are used for the calculation of weightings (Wikipedia 2012). They are computed and presented for each document (i.e. question paper). They are presented in a nut shell in Table 2.

The basic variables for term frequency computations are:

Number of concept terms occurring in one document = i;
Number of most frequently occurred word in one document = n;
Normalized term frequency, $tf = i/n$;_____(1)

The basic data for term frequencies are presented in Table 2.

Document No.	Total words	No. of most frequented word(n)	No. of concept terms appearing (i)			
			Activation	Demonstration	Application	Integration
Q1	187	15	11	5	6	1
Q2	206	25	6	7	8	2
Q3	184	16	9	12	6	2
Q4	173	21	6	10	8	2
Total no. of concept terms appearing			32	34	28	7

Table 2.Basic Data for Term Frequency

The total number of documents considered is N, which is 4.

The normalized term frequencies for each concept term of each document is presented in Table 3 (refer to equation 1).

Concept	Normalized term frequency				Total	Average
	Q1	Q2	Q3	Q4		
Activation	0.7333	0.2400	0.5625	0.2857	1.8215	0.4554
Demonstration	0.3333	0.2800	0.7500	0.4762	1.8395	0.4599
Application	0.4000	0.3200	0.3750	0.0952	1.1902	0.2976
Integration	0.0667	0.0800	0.1250	0.952	0.3669	0.0917

Table 3. Normalized Term Frequencies

The inverse document frequency is computed as:

$$idf = \log (N/K) ; \quad (2)$$

Where N = Total number of documents, which is 4 and
K: Number of occurrences of terms in all the documents considered (see total concept terms appearing in Table 2).

The final results are :-

Activation = -0.903089
Demonstration = -0.92959
Application = -0.845098
Integration = -0.243038

The negative sign is due to the fact that the total number of documents considered is less than the occurrences of terms in all the documents. The objective of this research work is to compare these frequency values of concept terms with respect to the benchmarks that were calculated based on manual methods. As the study is limited to comparisons, the idf is considered with modular values. Thus weighting would be

$$tf. \left| \log (N/K) \right| ; \quad (3)$$

Accordingly, the weighting for each concept term of each document is computed and presented in Table 4

Concept	Weighting of terms				Average
	Q1	Q2	Q3	Q4	
Activation	0.6622	0.2167	0.5080	0.2580	0.4113
Demonstration	0.3098	0.2603	0.6972	0.4427	0.4275
Application	0.3380	0.2704	0.3169	0.0805	0.2515
Integration	0.0162	0.0194	0.0304	0.2313	0.0223

Table 4. Weighting of portrayals in each document

The above values yield important conclusions.

8. CONCLUSIONS

The largest and the least present cognitive portrayals in every document as per benchmark values tally well with weighting of term frequency of all question paper documents considered. It is thus concluded that normalized term frequencies may be used for quantifying concept terms in individualized documents to an acceptable standards. However the benchmark values deviate from that of weighting term frequency on average values. This clearly shows that unlike quantification of keywords, quantifying concept words may not be reliable when inverse document frequency is computed on small number of documents. The work may be extended to large number of documents for further study.

REFERENCES

[1] Salton G, Buckley C (1988), "Term-weighting approaches in automatic text retrieval". Information Processing and Management 24 (5): 513–523, 1988.

[2] Sa-kwang Song, and Sung Hyon Myaeng, (2012), "A novel term weighting scheme based on discrimination power obtained from past retrieval results.", Information Processing and Management 48 (2012) 919–930, 2012.

[3] Masaru Ohba and Katsuhiko Gondow, (2005), "Toward Mining 'Concept Keywords' from Identifiers in Large Software Projects", ACM SIGSOFT Software Engineering Notes 30(4): 1-5, 2005.

[4] Merrill M.D., (2002), "First Principles of Instruction", Englewood Cliffs, NJ: Educational Technology Publications, 2002.

[5] Wikipedia (2012), http://en.wikipedia.org/wiki/Tf*idf, 2012

Authors

S.Florence Vijila is working as Assistant Professor in the Department t of Computer Science at CSI Ewart Women's Christian College,Tamil Nadu for the past 12 years.Her qualification is M.CA,M.Phil,Now she is doing Ph.D in the area of "Data Mining".Her work have been published in the International conference on "Innovations in contemporary IT research " proceedings.