Modified vector space model for protein retrieval computer science essay

This paper provides an sweetening of an bing method of information retrieval which is a alteration of Vector Space Model for information retrieval. This enhanced theoretical account is modified to be applied on protein sequence informations whereas the normal vector infinite theoretical account has been applied on text informations. The consequences show that the enhanced theoretical account achieved really good consequences in public presentation but the apparatus clip is someway high for a big aggregation of protein sequences

1. Introduction

The Vector Space Model (VSM) is a standard technique in Information Retrieval in which paperss are represented through the words that they contain. It was developed by Gerard Salton in the early 1960 ‘ s to avoid some of the information retrieval jobs. Vector spaces theoretical accounts convert texts into matrices and vectors, and so use matrix analysis techniques to happen the dealingss and cardinal characteristics in the papers aggregation.

It represents questions and paperss as vectors of footings which can be words from the papers or the question itself. The most of import thing is to stand for relevancy between paperss in this information infinite, which is achieved by happening the distance between the question and the papers [I] . The weight of relevancy of a question in the papers can be calculated utilizing some similarity steps such as cosine or dot merchandise or other measuring. Glenisson P. and Mathys J [4] have showed how the bag-of-words representation can be used successfully to stand for familial note and free-text information coming from different databases. They evaluated the

VSM by proving and quantifying its public presentation on a reasonably simple biological job.

They found that it can set up a powerful statistical text representation as a foundation for knowledge-based cistron look constellating [2] . In this work, we have modified the VSM technique to work with biological datasets. We used the papersfrequence (DF) alternatively of reverse papers frequence (IDF) . The consequences of the experiments show that the modified method give good consequences utilizing preciseness rating step.

2.

Vector Space Model

The VSM relies on three sets of computations. This theoretical account can work on selected index of words or on full text. The computations needed for the vector infinite theoretical account are: I. The weight of each indexed word across the fullpapers set demands to be calculated. This answers the inquiry of how of import the word is in the full aggregation. 2. The weight of every index word within a givenpapers (in the context of that papers merely) needs to be calculated for all N paperss. This answers the inquiry of how of import the word is within a individual papers.

3. For any question, the question vector is compared toevery one of the papers vectors. The consequences can be ranked. This answers the inquiry of which papers comes closest to the question, and ranks the others as to the intimacy of the tantrum.

The weight can be calculated utilizing this equation:

(“

Wi = tJi *logl21 dJi)Eql: where: if ; = term frequence (term counts) or figure of times a term I occurs in a papers. This accounts for local information. df, = papers frequence or figure of paperss incorporating term ID = figure of paperss in a database. The D/df, ratio is the chance of choosing a papers incorporating a queried term from a aggregation of paperss. This can be viewed as a planetary chance over the full aggregation. Therefore, the log (D/dfJ term is the86IJCSNS International Journal of Computer Science and Network Security, YOL. 7 No.

9, September 2007reverse papers frequence, IDF ; and histories for planetary information. 2. 1. VSM ExampleTo understand Eq I, allow ‘ s utilize a fiddling illustration. To simplify, allow ‘ s presume we cover with a basic term vector theoretical account in which we: I. Make non take into history WHERE the footings occur in paperss.

2. Use all footings, including really common footings and halt words. 3.

Make non cut down footings to root footings (stemming) . The undermentioned illustration [3] is one of the best illustrations on term vector computations available online. Suppose we query an IR system for the question “ gold silver truck ” The database aggregation consists of three paperss with the undermentioned content: D I: “ Cargo of gold damaged in a fire ” D2: “ Delivery of Ag arrived in a silver truck ” D3: “ Cargo of gold arrived in a truck ” Q: “ gold silver truck ” Vector infinite Model constructs the index tabular arraies as shown in Tables I and 2 by analysing the footings

of all paperss into words as in Table I and happen the frequence of each term in all paperss ; Table 2 does the same for the question. 2.

2 Similarity AnalysisThere are many different methods to mensurate how similar two paperss are, or how similar a papers is to a question in YSM. These methods include the: cosine, dot merchandise, Jaccard coefficient and Euclidian distance. In this paper we will utilize the cosine step which is the most common. The similarity step for the old illustration in subdivision 2. 1 can be calculated as follows: 1.

For each papers and question, compute all vector lengths (zero footings ignored)IDll = 00. 47712 + 0. 47712 + 0. 17612 + 0. 17612=~0.

5173 = 0. 7192 ID 21 = 00. 17612 + 0.

47712 + 0. 95422 + 0. 17612= A2001 = 1. 0955 ID31 = 00.

17612+ 0. 17612 + 0. 17612 + 0. 17612=~0. 1240 = 0. 3522: . ID, I = ~L.

. w2′. J I: .

IQI~) ~w~. J2. Calculate all point merchandises (nothing merchandises ignored) : Q-DI = 0.

1761 *0. 1761 = 0. 0310Q – D2 = 0. 4771 *0. 9542 + 0. 1761 *0. 1761 = 0.

4862 Q – D3 = 0. 1761 *0. 1761 + 0. 1761 *0.

1761 = 0. 0620. ‘ . Q-Di = LWQ wAA· .

. . J 1.

J13. Calculate the similarity values: Cosine vitamin E ~ Q -Dj Dj IQl*lDjl0. 0310 ~ 0. 08010.

5382 * 0. 7192Cosine vitamin E ~ Q – D 2 0. 4862 ~ 0.

8246D2 1 Q 1* I D2 1 0. 5382 *1. 0955Cosine vitamin E ~ Q-D3 _ 0. 0620 ~0. 3271D 3 1 Q 1 * 1 D 3 1 0. 5382 * 0. 3522, ‘ , Cosine 8D ; ~ SUll (Q, D ;) : Liter, vv Q, J, ” six, J, ‘ , SUll (Q, D ;) ~ ~L: .~, 2 ~L: w2Q.

J. ‘. JJ, IJCSNS International Journal of Computer Science and Network Security, VOL. 7 No. 9 September 2007ContributionWe can easy see from the old illustration that the normal VSM will non be suited for the protein sequence informations. This is because it uses the IDF in ciphering the weights, and as we saw in the illustration, IDF gives weight nothing if the term appears in all paperss and that is used for the halt or common words such as: a, an, the, of, ...

etc Since these words are really common they exist in all paperss, IDF gives these words rank 0 ; because normally the words that are in all paperss are non relevant. However, in protein there are no stop words as in text informations. So, the original method is non suited for protein informations because the being of a term in all protein sequences gives a significance and a weight must be given to this term. In this paper a little alteration on VSM is proposed to suit for protein sequence informations ; that is to utilize DF alternatively of IDF, where DF is the frequence of the term in all paperss (i. e. in how many paperss this term exists) .

This will give each papers its relevancy based on the frequence even if this term exists in all paperss so it will be suited for protein informations. We will utilize the cosine similarity step, which is the most common and has been proved by most research workers to give the best consequences for similarity [5] . 3. ExecutionWe have implemented the algorithm described in subdivision 2 in C scheduling linguistic communication. Experiments were run on a group of proteins that are known to be related. We tested the system on four protein households: ribosomal protein Ll, ribosomal protein L2, ribosomal protein L3, ribosomal protein L4, where each household has 50 proteins. 3. 1 Consequences and EvaluationThe plan has been tested on a aggregation of 200, 1000, 5000 and 10000 paperss, where the papers is a protein sequence as in Figure 1.

We have a file for protein sequences that we want to seek in, a file to input the question and an end product file that gives us the retrieved consequences. The trial of the plan has been applied as follows: We chose a sequence of aminic acids as a question from the aggregation of protein sequence, for illustration from LI, and fit it with the whole file and see the consequences. The relevant paperss would be those from LI, because we get the question from Ll. 87& gt ; NFOI724288 Ribosomal protein L3 [Desulfovibrio vulgaris] MAEKMGILGRKIGVTRIF ASDGSA VA VTVIK AGPCPVTQVKTV ATDGYDAIQIAFDEAKEKH LNKPEIGHLAKAGKGLFRTLREIRLEAP AA YE VGSELDVTLFATGDRVKVSGTSIGKGYQGVM RRWNF AGSKDTHGCEKVHRSGGSIGNNTFPGFigure I: one protein sequence [6]3.

2 EvaluationWe used the standard IR rating to measure the algorithm. The preciseness gives the metric per centum of the figure of relevant paperss

retrieved to the paperss retrieved. Number of relevant paperss retrievedPreciseness = Number of retrieved paperssThis step gives us how accurate the method is from the figure of relevant paperss we retrieved. If the preciseness = I, this implies that the algorithm has successfully identified all relevant paperss. Using this method on 200 paperss with 50 relevant paperss and utilizing 10, 20 and 50 as the question length, we get the undermentioned consequences: We can see from Tables 3- 5 that the preciseness for a question of length 10 is 80 % this is because the question length is non long plenty and can be found in many protein sequences, whereas for a question of length 20AA50, the preciseness is 100 % for a cutoff = 5-10, and make 39 % for a cutoff = 100, and this is good consequences for preciseness step. 3. 3. Apparatus TimeThe apparatus clip is the clip for building the index tabular arraies showed in Tables I and 2 in add-on to the put to deathing clip of the plan starting from come ining the question inquiring for the retrieved paperss until it gives the retrieved paperss.

To cipher the apparatus clip of the plan, a aggregation of 200, 1000, 5000 and 10000 paperss has been used, taking into history that this apparatus clip is for the first tally which includes the constructing of the indexes. We can see from Table 6 that the apparatus clip is rather sensible for little paperss up to 5000, but after that the apparatus clip increases quickly. This can be improved by parallelizing the plan administering the informations on multiple nodes which will diminish the apparatus clip. IJCSNS International Journal of Computer Science and Network Security, VOL.

7 No. 9 September 2007

4. Decision

In this paper a modified theoretical account of VSM which is applied on protein sequences information has been introduced. The modified method achieved good consequences and good public presentation for recovering the protein informations.

Using the preciseness rating step, it gives a preciseness of I for a cutoff = 10, and 0. 39 for a cutoff of 100. The consequences show that for a little papers aggregation the apparatus clip is sensible, but for big aggregation it gives really large apparatus clip. Our following measure is to prove the plan on larger informations and compare the public presentation of this modified method with other similar methods. We besides intend to research the application of analogue techniques to cut down the big apparatus clip.