Safa'a I. Hajeer / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue4, July-August 2012, pp.2085-2090

Vector Space Model: Comparison Between Euclidean Distance & Cosine Measure On Arabic Documents

Safa'a I. Hajeer
Department of Computer Information Systems

Abstract

The purpose of this research is to give an idea about Euclidean distance and cosine measure based on Arabic documents collection, and gives the comparison points between those measures.

The most common points to compare are the system performance with these two measures by give the attention on time, space and recall/precision evaluation measures. The time measure represents the time needed to retrieve the relevant documents to specific query. On the other hand, space represents the capacity of memory needed for reach the results. Then at last find the recall/precision the saw the effectiveness of the system.

A collection of 242 Arabic Abstracts from the proceeding of the Saudi Arabian National Computer Conferences, and tested the system with a lot of sample queries to see that its work go correctly. Note: the index terms used in full words after removing the stop words, to reach better results by the exact matches.

The results have proved that the Euclidean distance had exact accuracy in compared the cosine measure that theoretically exact, but suffers from rounding errors. However, the comparison of time complexity of the two measures we found them the same define as O(N). Note the time for calculate the cosine measure equation need more time and space than the Euclidean distance although the complexity time is the same for both.

Keywords: IR, Vector space model, ranking algorithm, Euclidean distance, Cosine measure.

Introduction

Information Retrieval (IR) is a field devoted primarily to efficient, automated indexing and retrieval of documents. There are a variety of sophisticated techniques for quickly searching documents with little or no human intervention. [6]
Traditional information retrieval systems usually adopt index terms to index and retrieve documents. An index term is a keyword (or group of related words) which has some meaning of its own (i.e. which usually has the semantics of the noun). [1] That's mean in a simple way an words which appears in the text of a document in a collection and its fundamental for information retrieval task to help the users to find the information which they need in an easy way.

So the information retrieval systems issue is predicting which documents are relevant and which are not. Such a decision is usually dependent on a ranking algorithm which attempts to establish a simple order of the document retrieved. Documents appearing at the top of this ordering are considered to be more likely to be relevant.

A ranking algorithm operates according to basic premises (evidences, principle, ideas, foundations, grounds) regarding the nation of document relevance. Distinct sets of premises yield distinct information retrieval models. One of the IR models is vector space model.

The Vector Space Model (VSM) is a popular to Information retrieval system implementation which it based on the idea of represented documents by vectors (arrays of numbers) in a high-dimensionality vector space. To look deeply in the vector space model read the next section.

Vector Space Model

The vector model defines a vector that represents each document, and a vector that represents the query [9]. There is one component in each vector for every distinct term that occurs in the document collection. Once the vectors are constructed, the distance between the vectors, or the size of the angle between the vectors, is used to compute a similarity coefficient [2].

Consider a document collection with only two distinct terms, a and B. All vectors contain only two components. The first component represents occurrences of a, and the second represents occurrences of B. The simplest means of constructing a vector is to place a one in the corresponding vector component if the term appears, and a zero, if the term does not appear. Consider a document D1, that contains two occurrences of term a and zero occurrences of term b. The vector, represents this document using a binary representation. This binary representation can be used to produce a similarity coefficient, but it does not take into account the frequency of a term within a document. By extending the representation to include a count of the number of occurrences of the terms in each component, these frequencies can be considered [2].
Early work in the field used manually assigned weights. Similarity coefficients that employed automatically assigned weights were compared to manually assigned weights. Repeatedly, it was shown

that automatically assigned weights would perform at least as well as manually assigned weights [9].

Euclidean Distance

Around 300 BC, the Greek mathematician Euclid laid down the rules of what has now come to be called "Euclidean geometry", which is the study of the relationships between angles and distances in space. [4]

Euclidean distance is a measure to find the distance between the query and the document.

The reader can see the formula of Euclidean distance later on (in the section of The Algorithm).

Cosine Measure

Cosine measure: one of the similarity measures of VSM which represents the documents is closest to the query (user request).

Thus a cosine value of zero meant that the query and document vector were orthogonal to each other and meant that there was no match or the term simply did not exist in the document being considered.

The reader can see the formula of Cosine Measure later on (in the section of the Algorithm).

The Algorithm:

1- The frequency of each term is found and a preliminary document vector is formed.

2- The document vectors are normalized using the below formula:

co-ordinate of the vector is = (term 1 frequency) / sqrt [(term 1 freq)^2 + (term 2 freq)^2 +.....+ (term n freq)^2]

3- Similarly all the other co-ordinates of the document vector and query (Note: The query is also treaded as a document) are calculated. Note that the co-ordinate= weight of the terms.

4- Find the Euclidean distance

The Euclidean distance formula is used then to calculate the distance between the query and the document.

$$E.D\ (D, Q) = \text{sqrt}\ [\sum (t_k - q_k)^2]$$

5- find the Cosine measure

The Cosine Measure formula is used to calculate the cosine measure between the query and the document.

$$C.M(D,Q) = \sum(t_k q_k) \ / \ [\ \text{sqrt}(\ \sum(t_k)^2\ )\ *\ \text{sqrt}(\ \sum(q_k)^2\ )\ ]$$

6- The results are then ranked. For the Euclidean Distance (E.D) ranking is done from lowest distance to highest distance (i.e. lowest E.D are placed first).

For the Cosine Measure (C.M) ranking is done from highest value to the lowest value. (i.e. highest C.M are placed first).

The reason for this: If the angle between the vectors is small they are said to be near each other and a small angel means a high cosine value (for Ex: cos 0^o = 1)

7- At last check the effectiveness of the system by measuring the Accuracy (recall/precision).

Measuring Accuracy

Accuracy of an Information Retrieval System is commonly measured using the metrics of precision and recall. These are defined in Equation one below. Precision is the measure of how much junk (non-relevant) documents get returned for each relevant one. Recall is the measure of how many of the relevant documents were found no matter what else was found. These measures assume a prior knowledge of which documents are relevant to each query. [2]

Precision= number of relevant document retrieved/ Total # of retrieved document

Recall=number of relevant document retrieved/ Total # of relevant document

Equation1: Precision and recall Definition

Results

We can see the results that reached from the comparison:

Measure Name	Description	Data Structure	Accuracy(Theoretical)	Theoretical time complexity
Cosine Measure	A simple brute force algorithm that calculates the Cosine measure between the query profile and all profiles in the system, and then returns a list of the best matching ones documents. For the cosine measure to work all profiles must be unit length, and therefore they must be normalized first. This can be done as a one time preprocessing step so it does not impose too severe an overhead. Unfortunately reduction does not preserve unit length and therefore	Linked list of all document profiles	Theoretically exact, but suffers from rounding errors, because of excessive normalizing steps	O(N)

	every time a profile is reduced, it needs to be normalized again. These repeated normalization operations cause computational costs and reduce the accuracy of this method.			
Euclidean distance	A very simple brute force algorithm that calculates the Euclidean distance from the query profile to all profiles in the system, and then returns a list of the best matching ones.	Linked list of all document profiles	Exact	O(N)

To Focus at Accuracy in this information retrieval system, see the graph:

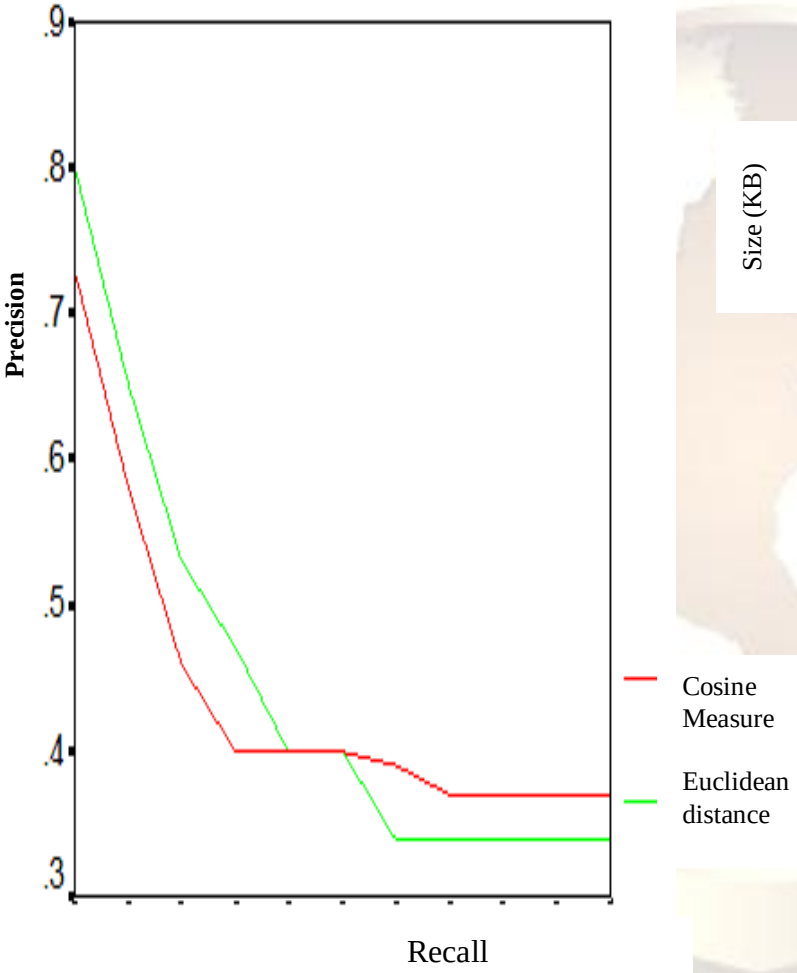

Figure 1: Recall\precision using 50 Query as a sample

To know more about the space which is used by the system during the testing at the documents - see below graph

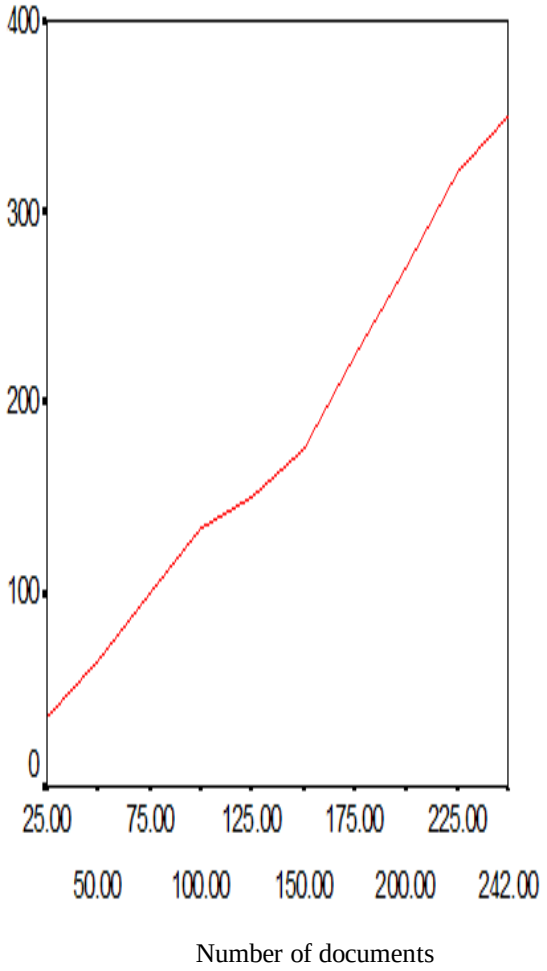

Figure 2: space requirement

Conclusion

Through the theoretical analysis and experimental results: for the normalized data, Euclidean distance and cosine measure becomes even more similar.

The cosine measure suffers from rounding error because of excessive normalizing steps. Note the repeated of normalization operations cause computational costs, so the looking for a normalized data without any need to do normalization operation on them to reach the best results without any rounding error.

In future work, the plan to extend this project to analyze other distance measures, such as Manhattan distance, Chebishov distance, power distance …etc.

References:

[1] Baeza-Yates Ricardo, Ribeiro-Neto Berthier (1999), *Modern Information Retrieval*, New York, USA.
[2] Chowdhury Abdur, McCabe M.Catherine (1993), *Improving Information Retrieval Systems using Part of Speech Tagging*.
[3] Deitel (2003), *Java how to program*, USA, fifth edition.
[4] Euclidean Space – Wikipedia (2012), the free encyclopedia, *http://en.wikipedia.org/wiki/Euclidean_space*, available on: July 2012.
[5] G. Salton, M. J. McGill (1983), *introduction to Modern Information Retrieval*, McGraw-Hill Inc.
[6] Information Retrieval & computational Geometry (2012), www.ddj.com/dept/architect/184405928, available on: July 2012.
[7] Information retrieval – Wikipedia (2012), the free encyclopedia, *http://en.wikipedia.org/wiki/Information_retrieval*, available on: Jan. 2012.
[8] Qian Gang, Sural Shamik, Gu Yuelong, Pramanik Sakti (2004), *Similarity between Euclidean and cosine angle distance for nearest neighbor queries*, to appear in Proceedings of the 19th Annual ACM Symposium on. Applied Computing (SAC 2004), Nicosia, Cyprus.
[9] Salto, G., Wong, A., and C.S. Yang (1975), *A Vector Space Model for Automatic Indexing*, Communication of ACM, New York, USA, Vol. 18 Issue 11, p.p. 613- 620.

Appendix A

The following are the queries used to test the system:

Q1:استخدام الحاسب الالي
Q2:استرجاع المعلومات
Q3:الادارة و التخطيط
Q4:التدريب و التعليم
Q5:الترميز و التشفير
Q6:التعليم بمساعدة الحاسب
Q7:التعليم بواسطة الحاسب
Q8:الحاسب الالي
Q9:الحاسبات الصغيرة
Q10:الحاسبات المتناهية الصغر
Q11:الحاسوب و التعليم
Q12:الحج و العمرة
Q13:الحرف العربي
Q14:الخطة الوطنية للمعلومات
Q15:الخليج العربي
Q16:الدوائر المتكاملة
Q17:الذكاء الاصطناعي
Q18:الذكاء الالي
Q19:العالم العربي
Q20:القران الكريم
Q21:الكلمات العربية
Q22:اللغات الطبيعية
Q23:اللغة العربية
Q24:المدرسة الالكترونية
Q25:المملكة العربية السعودية
Q26:الموارد البشرية
Q27:النص العربي

Q28:امن المعلومات
Q29:انظمة الحاسبات الالية
Q30:برامج الحاسب الالي
Q31:برمجة الحاسبات الالية
Q32:بنوك المعلومات
Q33:تدريس مواد الحاسب
Q34:تركيب الجملة العربية
Q35:تطبيقات الكمبيوتر
Q36:تعريب البرامج
Q37:تعريب الحاسبات الالية
Q38:تعريب الحاسوب
Q39:تعليم الكمبيوتر
Q40:تقنية المعلومات
Q41:تمييز الاشكال بواسطة الحاسب الالي
Q42:جامعة الملك سعود
Q43:جامعة الملك عبد العزيز
Q44:جمعية الحاسبات السعودية
Q45:شبكات الحاسب الالي
Q46:شبكة اتصالات الحاسبات
Q47:شبكة الاتصالات
Q48:علوم الحاسب و المعلومات
Q49:قواعد البيانات
Q50:قواعد المعلومات

Appendix B

The following are a sample of documents from the collection used to test the system:

رقم 32
صنف بنوك المعلومات
نوع مؤتمر
عنو تحليل العلاقة بين الموارد البشرية ونظم المعلومات
مؤل الصعيدي , ابراهيم احمد
جهه قسم المحاسبة , كلية العلوم الادارية والسياسية , عين شمس , مصر
عنم المؤتمر الوطني العاشر للحاسب الآلي , مركز الحاسب الآلي - جامعة الملك عبدالعزيز
مجل 2
صفح 916 - 950
نشر 1408 هـ
ناش مركز النشر العلمي, جامعة الملك عبدالعزيز, جده
لغه العربية
ملخ يعتبر الانسان هو المحرك الرئيسي لاي نشاط مهما بلغت درجة التقدم الحضاري والتكنولوجي المعاصر , ولا يجوز باي منطق ان نقارن قدرات الانسان بقدرات اي نظام او جهاز لان الانسان يمتلك العقل الذي ميزه به الله على سائر الكائنات " الم نجعل له عينين ولسانا وشفتين وهديناه النجدين " - فالانسان بذكائه الفطري هو الذي اكتشف نظام الارقام الحسابية , والعمليات الجبيبة المختلفة , وفي استطاعة التذكر والاستدلال والتحليل , وتخزين البيانات في ذاكرته , واتخاذ القرارات ورسم السياسات كما في امكانه الاحتفاظ بالبيانات والمعلومات في وسائل خارجية في صورة ملفات , او اشرطة او اجهزة الحاسب والميكروفيلم , بل فقد دونها في بداية حياته على الجدران كما فعل القدماء المصريون . ان هذه الامور تعتبر من البديهيات لذلك فلا عجب ولا غرابة ان تهتم النظم الادارية الحديثة " بالانسان - كانسان " وتوفر له جميع الامكانات للابتكار والتجديد وتهيىء له الجو النفسي الذي يساعده على الانتاجية والخلق والتجديد مما دفع المحاسبين بدورهم بالمطالبة بتقدير قيم - الموارد البشرية العاملة في المشروعات والنظم وادراجها كأصل من الاصول في قائمة المركز المالي بالرغم من الصعوبات التي تكتنف تلك العملية. ويتناول هذا البحث توضيح أهمية الدور الذي تلعبه الموارد البشرية كاحد عناصر نظم المعلومات من خلال التعرض للموضوعات التالية : الاتجاهات المعاصرة لتقييم الموارد البشرية , الموارد البشرية كاحد عناصر نظم المعلومات الادارية , التغيرات السلوكية للموارد البشرية نتيجة ادخال النظم الالكترونية للمعلومات .
مطب نسخ ورقية .

رقم 40
صنف البرمجة
نوع مؤتمر
عنو تقليل تكاليف بناء الطرق باستعمال البرمجة الديناميكية
مؤل عامر , رشدي عبدالرحمن
جهه قسم الحاسب الآلي , جامعة الملك عبدالعزيز , جده
عنم المؤتمر الوطني العاشر للحاسب الآلي , مركز الحاسب الآلي - جامعة الملك عبدالعزيز
مجل 2
صفح 619 - 628
نشر 1408 هـ
ناش مركز النشر العلمي , جامعة الملك عبدالعزيز , جده
لغه الانجليزية
لغة العربية
ملخ يقدم البحث نموذجا يعتمد على اسلوب البرامج الديناميكية للحصول على التصميم الامثل لخط من خطوط المرافق مثل الطرق او خطوط السكك الحديدية او القنوات . ويهدف النموذج الى تقليل تكلفة الحفر والردم اللازمين لتسوية المناسيب الطبيعية لسطح الارض على طول المسار . هذا مع تحقيق متطلبات التصميم من حيث الانحدار والانحناء لكل مرحلة من مراحل الطريق . ويسمح النموذج بتغير هذه المتطلبات على طول الخط كما يسمح بتحقيق متطلبات جانبية تتعلق باطوال اجزاء الخط او موازنة احجام الحفر والردم . ويقبل النموذج التغير في تكلفة الحفر والردم حسب تغير طبيعة التربة على طول المسار . كما انه في حالة الاعماق الكبيرة للحفر او يمكنه اعتبار حلول بديلة مثل الانفاق بدلا من حفر او ردم الجسور ويتم الحل على مرحلتين . في المرحلة الاولى تخزن القرارات المثلى . والتكاليف المناظرة في هيكل للبيانات على شكل شجره , وفي المرحلة الثانية يستخرج الحل الامثل بعبور الشجرة من الطرف الامثل الى الجذر . مطب نسخ ورقية

رقم 44
صنف الحاسبات الآلية - لغات
نوع مؤتمر
عنو تخطيط وتصميم شبكات اتصالات الحاسبات نظرة خاصة بالمملكة العربية السعودية
مؤل غنيمي , محمد اديب رياض , شاهين , حسين اسماعيل , نور , يوسف محمد عجب
جهه كلية الهندسة , جامعة الملك عبدالعزيز , جده
عنم سجل بحوث المؤتمر والمعرض الوطني السابع للحاسبات الالكترونية
صفح 10
نشر 1404 هـ
ناش معهد الادارة العامة , السعودية
لغه العربية
ملخ في هذه المقالة نعرض الطرق المختلفة لتخطيط وتصميم شبكات من منظور المملكة العربية السعودية , وتركز المقالة على مشكلة المهام المحددة