Comparison on the Effectiveness of Different Statistical Similarity Measures

Safaa I. Hajeer
Department of Computer
Information Systems

ABSTRACT

Document retrieval is the process of matching of some sated user query against a set of free-text records (documents), its one major technique for organizing and managing information. This project was concerned with studying which of the different statistical measures in IR have the most effectiveness on document retrieval using a unified set of documents. The results show that the Cosine Similarity Measure is the best of other seven measures (Inner Product, Dice Coefficient, Jaccard Coefficient, Inclusion Similarity Coefficient, Overlap Coefficient Measure, Euclidean distance Measure and Manhattan Distance Measure (City Block Distance) for both languages, with precision on Arabic collection 38% and recall 53.2%. On English collection, the precision is 25% and recall 65%.

KEYWORDS

Information Retrieval (IR), Vector space model, ranking algorithm, Similarity Measures.

1. INTRODUCTION

For thousands of years people have realized the importance of archiving and finding information. With advent of computers, it become possible to store large amounts of information, and finding useful information from such collections became a necessity. The field of Information Retrieval (IR) was born in the 1950s out of this necessity [1].

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. [2]

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). [3] 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 both the query and each document as vectors in the term space. A measure of the similarity between the two vectors is computed [4]. Similarity measures take advantage of richer representation (e.g. term weights based on occurrence frequency). Most similarity measures can be classified into one of four types: (1) Angular measures (e.g. the cosine measure), (2) Distance measures (e.g. Euclidean distance), (3) Association coefficients (e.g. Jaccard coefficient), (4) Probabilistic measures.

Most studies found in the text mining literature use different document sets, which make it difficult to determine the best text classification method [5]. This study was concerned with studying the efficiency of the first three similarity types by using several methods belong to each type using a unified set of text documents.

2. RESEARCH BACKGROUND

Vester et al. said in [6], Information retrieval dates more than 4000 years back to the beginning of written language, as information retrieval is related to knowledge stored in textual form. Today text has grown to become:

> "… The primary way that human knowledge is stored, and after speech, the primary way it is transmitted."

Traditionally, information retrieval was a manual process, mostly happening in the form of book lists in libraries, and in the books themselves, as tables of contents, other indices …etc. these lists/tables usually contained a small number of index terms (e.g. title, author and perhaps a few subject headings) due to the tedious work of manually building and maintaining these indices.

The above was true through most of history up until the middle of the 20th century, where the digital computer fundamentally changed the way that person was able to store, search and retrieve textual information as noted in [6]. As a result, IR has grown well beyond its previous limited form, mostly concerned with indexing and searching books and other kinds of textual information.

Today, information retrieval plays a much larger part of our lives- especially with the advent of Internet, and the World Wide Web (the Web) in particular. In [6], during the last ten years, the amount of information available in electronic form whether on documents in computers or on the Web has grown exponentially. Almost any kind of desired information is available, including: news and message files, software libraries, multimedia repositories, commercial information, Bibliographic collections, encyclopedias …etc.

Furthermore, the amount of documents managed in organizational intranets that represent the accumulated knowledge of the organizations is also quickly growing, and efficient access to these documents has become vital to the success of modern organizations.

In [6], the writer said, Information retrieval is at the center stage of this "revolution" and is necessary condition for its continuing expansion into even more areas of our lives. However, technologies enter many areas of our life, people still find it difficult (if not impossible) to consistently locate and retrieve information relevant to their needs. As Roussinov and Chen pessimistically put it:

> "Our productivity in generating information has exceeded our ability to process it, and the dream of creating an information- rich society has become a nightmare of information overload."

In modern information retrieval systems, several models exist to represent the information contained in a large collection of textual documents. Vector Space Model is one of these models, the first proposed it by Salton at el. In the paper "A vector Space Model for Automatic Indexing" [7] [11]. In the vector space model, a document and a query are represented as two term vectors in a high-dimensional term space. Each term is assigned a weight that reflects its "importance" to the document or the query. Given a query, the relevance status value of a document is given by the similarity between the query vector and document vector as measured by some vector similarity measure, such as the cosine of the angle formed by the two vectors, Euclidean distance which is study the relationships between angels as reported in [8]. 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 according to [9]. One key question in document retrieval is how to rank documents based on their degrees of relevance to a query. Much effort has been placed on the development of ranking functions; some of them based idea on different similarity measures, but the comparison of their effectiveness on the unified set of data is still in studying especially for Arabic documents.

Note: Early work in the field of Vector Space Model (VSM) 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 as stated in [10] [11] but it's faster, easier and more precise.

3. EVALUATION

In order to evaluate the classification system, a collection of English documents collected from NPL (national physics Laboratory in the United Kingdom (UK). The collection is available on the net. This collection contains 800 documents with different size, and tested the system with 20 queries as a sample, the queries represented in Appendix A and a sample of documents from the collection in Appendix B. the system uses English stop word list contains 319 words.

Another collection is used, a collection of 240 Arabic abstracts from the proceeding of the Saudi Arabian National Computer Conferences. The system tested with 20 queries as a sample. The queries represented in Appendix A and a sample of documents from the collection in Appendix B. the system uses Arabic stop word list contains 1459 words.

On English collection (NPL dataset), **Table 1** represent the result accuracy of the system. The results indicate that the Cosine similarity measure is the most statistical measure as compared to the other seven measures, with 25% precision and 65% recall, then the Inner Product become next with 23.6% precision and 64.3% recall. The third measure is Inclusion Similarity Coefficient with 21.2% precision and 63.5% recall. The fourth one is Overlap Coefficient Measure with 20.7% precision and 62.7% recall. The fifth one is Jaccard Coefficient with 20.4% precision and 61.8% recall. The sixth is Euclidean distance with 20% precision and 60% recall. The seventh one is Dice Coefficient with 18% precision and 57.1% recall. The last one is Manhattan Distance Measure (City Block Distance) with 50% recall and 13.6% precision.

Table 1: The accuracy measurement of the System on English documents.

Statistical Measure	Recall	Precision	F-Measure
Inner Product (dot Product)	0.643	0.236	0.345
Dice Coefficient	0.571	0.180	0.274
Jaccard Coefficient	0.618	0.204	0.307
Inclusion Similarity Coefficient	0.635	0.212	0.318
Cosine Similarity Measure	0.65	0.250	0.361
Overlap Coefficient Measure	0.627	0.207	0.311
Euclidean distance Measure	0.600	0.200	0.399
Manhattan Distance Measure (City Block Distance)	0.500	0.136	0.214

Figure 1- 3 reflect the results on **Table 1** which shows that Cosine Similarity Measure is the best measure to use.

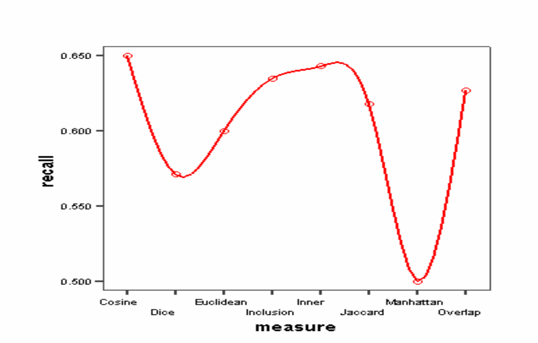


Fig. 1: Recall of Eight Measures on English Collection.

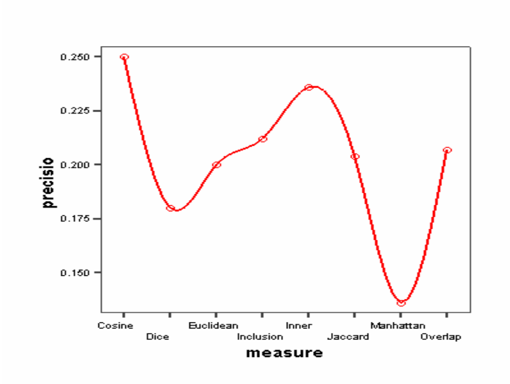


Fig. 2: Precision of Eight Measures on English Collection.

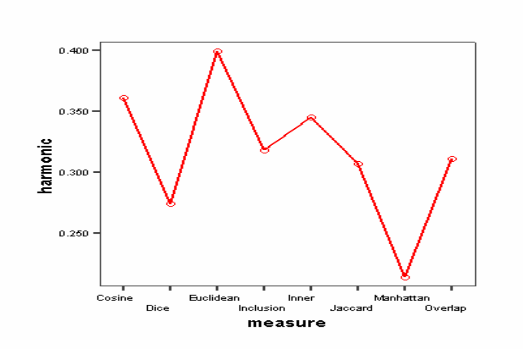


Fig. 3: Harmonic Measure of Eight Measure on English Collection.

On Arabic collection, **Table 2** and through **Figures** represent the result accuracy of the system. The results indicate that the Cosine similarity measure is the most effective one as compared to the other seven measure, with 38% precision and 53.2% recall, then the Inner Product become next with 37.6% precision and 53% recall. The third measure is Inclusion Similarity Coefficient with 37.4% precision and 52.8% recall.

The fourth one is Overlap Coefficient Measure with 37.2% precision and 52.7% recall. The fifth one is Jaccard Coefficient with precision 36% and 52.6% recall. The sixth is Euclidean distance with 35.7% precision and 52.5% recall. The seventh one is Dice Coefficient with precision 35.6% and 52.2% recall. The last one is Manhattan Distance Measure (City Block Distance) with 51.8% recall and 15.6% precision.

Table 2: The accuracy measurement of the System on Arabic documents.

Statistical Measure	Recall	Precision	F-Measure
Inner Product (dot Product)	0.530	0.376	0.440
Dice Coefficient	0.522	0.356	0.423
Jaccard Coefficient	0.526	0.360	0.427
Inclusion Similarity Coefficient	0.528	0.374	0.438
Cosine Similarity Measure	0.532	0.380	0.443
Overlap Coefficient Measure	0.527	0.372	0.436
Euclidean distance Measure	0.525	0.357	0.425
Manhattan Distance Measure (City Block Distance)	0.518	0.156	0.240

Figure 4- 6 reflect the results on Table 2 which shows that Cosine Similarity Measure is the best measure to use.

The final result shows from this research that Cosine Similarity Measure is the most effectiveness measure for both Arabic and English languages. This result learnt from the comparison of it with the other seven measures used by the system, so we encourage other IR researchers, if they want to select one of the above measures, they can select the cosine measure to make their search system more efficient

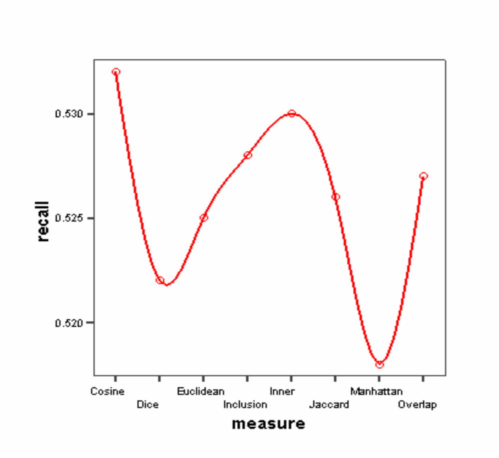


Fig. 4: Recall of Eight Measures on Arabic Collection.

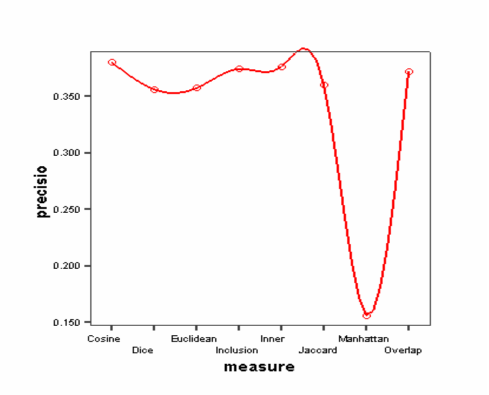


Fig. 5: Precision of Eight Measures on Arabic Collection.

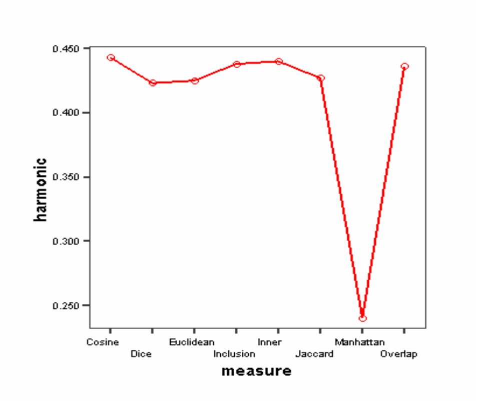


Fig. 6: Harmonic Measure of Eight Measure on Arabic Collection.

4. CONCLUSION

In this research, Several Statistical Measures is employed to construct a retrieval model for English and Arabic documents. This research address two issues; Studying the effectiveness of each different statistical measure on a unified set of documents, and Comparing the effectiveness of different statistical measures in order to find the most effective one to classify documents.

The results show that the Cosine Similarity Measure is the best of other seven measures (Inner Product, Dice Coefficient, Jaccard Coefficient, Inclusion Similarity Coefficient, Overlap Coefficient Measure, Euclidean distance Measure and Manhattan Distance Measure (City Block Distance) for both languages, with precision on Arabic collection 38% and 53.2% recall. On English collection, the precision is 25% and 65% recall. So according to this research until the moment of writing is the Cosine Similarity Measure is the best to use in ranked retrieval system for a given queries.

According to these experimental results the Cosine Similarity measure is the best for Document retrieval technique.

5. FUTURE WORK

In future work, the plan is to extend this project to test other measures, such as Lucene's Similarity measure, Chebyshev distance, Sokal and Sneath similarity measure, Bray Curit (Sorensen) distance, spearman distance and power distance. Moreover add a stemmer to the system to compare results of it with and without stemmer.

We plan to study the semantic of both documents and user query via combine these statistical measures to Conceptual Graph (CG) in another version of this system called AutoClassify2. Conceptual Graph (CG) is one of many natural language processing (NLP) techniques, which concern in the semantics of documents and query proposed. We think this will improve the quality of retrieved documents and take more acceptance and trusty of users who looking to find the best answer for their query in short time.

Comparing the effectiveness of statistical measures used by the system on another languages collection, like Turkish, Spanish, Italian …etc. Surely after learning their nouns and stop words.

6. REFERENCES

[1] Singhal A. (2001), *Modern Information Retrieval: A Brief Overview*, IEEE Data Engineering Bulletin, Vol. 24, No. 4, pp. 35-43.

[2] Stephens R. (2004), *Information Retrieval & computational Geometry*, www.ddj.com/dept/architect/184405928, available on October, 2008.

[3] Baeza-Yates R. and Ribeiro-Neto B. (1999), *Modern Information Retrieval*, ACM Press, New York.

[4] Grossman D. and Frieder O. (2004), *Information Retrieval Algorithms and heuristics*, Netherlands, USA.

[5] Al-Sinjilawi S.and Al- Kabi M. (2007), *A comparative study of efficiency of different measures to classify Arabic text*, University of Sharjah of pure & Applied sciences, Vol. 4, No. 2.

[6] Vester K. and Martiny M. (2005), *Information Retrieval in Document spaces using clustering*, Master Thesis, Technical University of Denmark, Denmark.

[7] Garcia E. (2008), *Understanding Inverse Document Frequency (IDF)*, IR Watch Newsletter, USA.

[8] Zhai C. (2007), *A Brief Review of Information Retrieval Models*, University of Illinois at Urbana Champaign, USA.

[9] *Euclidean Space – Wikipedia, the free encyclopedia*, http://en.wikipedia.org/wiki/Euclidean_space, available on October, 2008.

[10] Chowdhury A. (2001), *On the Design of reliable efficient information systems*, Thesis for Doctor of philosophy in Computer Science, Illinois Institute of Technology, Chicago, USA.

[11] Salton G., Wang A. and Yang C. (1975), *A Vector Space Model for Automatic Indexing*, Communication of the ACM, Vol. 18, No. 11, pp. 613-620.

Appendix A

Tested Queries

The following are the queries used to test the system:

- For English documents:

Q1: Secondary emission of electrons.

Q2: spherical harmonic analsis of the earths magnetic field

Q3: solar flares

Q4: equations governing the propagtion of electronic and hydromanetic waves in the solar corona

Q5: magnetic storms

Q6: estimates of the density of ionization and temperature in the solar corona

Q7: Integral spectrum

Q8: ferromagnetic techniques for computer stores

Q9: magnetic film memory

Q10: circuitry capable of generating extremely narrow pulses

Q11: parametric amplifiers

Q12: variable capacitance amplifiers

Q13: advantages of parametric amplifiers

Q14: oscillators

Q15: observations of rapid fluctuations in the earths magnatic field and thier relation to the propagation of hydromagnetic waves

Q16: synthesis of filters

Q17: diurnal variations of fluctuations in the earths magnetic field

Q18: measurments of ionospheric drifts near the equator

Q19: cosmic ray results

Q20: beam of charged particles

- For Arabic documents:

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: انظمة الحاسبات الالية

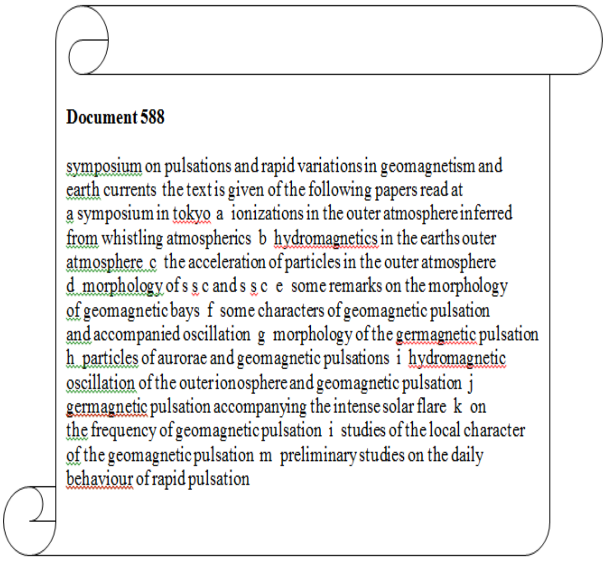
Q20: برامج الحاسب الالي

Appendix B

Sample of Documents

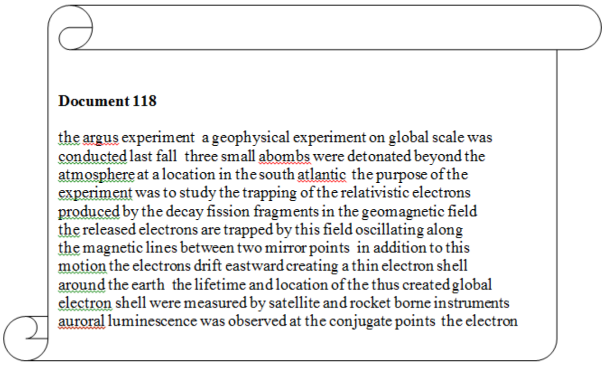
The following are a sample of documents used in this research:

- For English:



Document 588

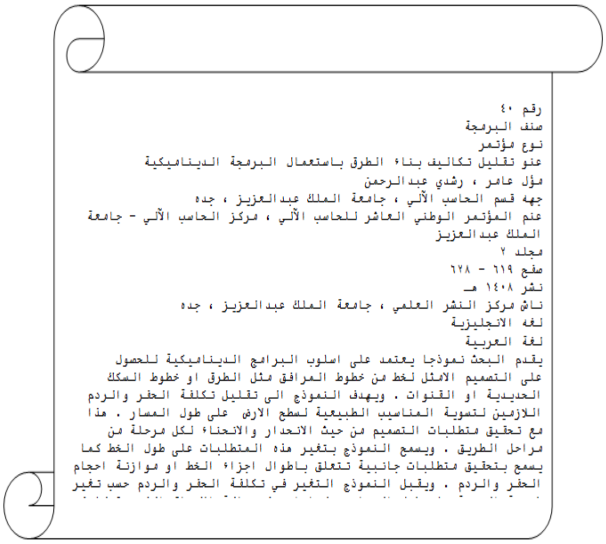
symposium on pulsations and rapid variations in geomagnetism and
earth currents the text is given of the following papers read at
a symposium in tokyo a ionizations in the outer atmosphere inferred
from whistling atmospherics b hydromagnetics in the earths outer
atmosphere c the acceleration of particles in the outer atmosphere
d morphology of s s c and s s c e some remarks on the morphology
of geomagnetic bays f some characters of geomagnetic pulsation
and accompanied oscillation g morphology of the germagnetic pulsation
h particles of aurorae and geomagnetic pulsations i hydromagnetic
oscillation of the outer ionosphere and geomagnetic pulsation j
germagnetic pulsation accompanying the intense solar flare k on
the frequency of geomagnetic pulsation i studies of the local character
of the geomagnetic pulsation m preliminary studies on the daily
behaviour of rapid pulsation



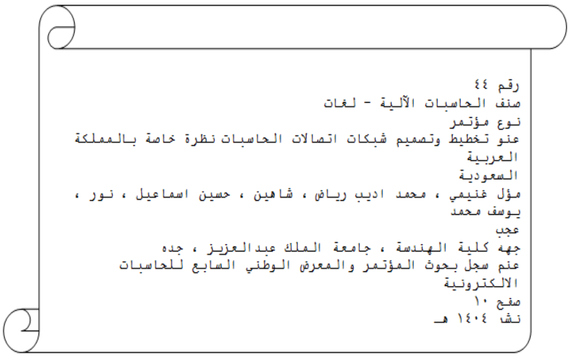
Document 118

the argus experiment a geophysical experiment on global scale was
conducted last fall three small abombs were detonated beyond the
atmosphere at a location in the south atlantic the purpose of the
experiment was to study the trapping of the relativistic electrons
produced by the decay fission fragments in the geomagnetic field
the released electrons are trapped by this field oscillating along
the magnetic lines between two mirror points in addition to this
motion the electrons drift eastward creating a thin electron shell
around the earth the lifetime and location of the thus created global
electron shell were measured by satellite and rocket borne instruments
auroral luminescence was observed at the conjugate points the electron

- For Arabic:



رقم ٤٠
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
مجلد ٢
صفح ٦١٩ - ٦٢٨
نشر ١٤٠٨ هـ
ناش مركز النشر العلمي ، جامعة الملك عبدالعزيز ، جده
لغه الانجليزية
لغة العربية
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



رقم ٤٤
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 ١٠
نش ١٤٠٤ هـ