# The measurement of similarity in stock data documents collections

Nikitas N. Karanikolas[1]

[1] Department of Informatics, Technological Educational Institute of Athens, Athens, Greece
Tel: +302105385882, Fax: +302105910975, E-mail: nnk@teiath.gr

In this paper we focused on the similarity of documents, stored in a documents collection, against a submitted query. This topic is also known as Document versus Query similarity or Nearest Neighbors method. The analysis of experimental queries conducted to a collection of 4984 financial documents (related to stock data and varying in size from few words to a lot of pages) give us an indication that the Nearest Neighbor method presents a "preference" to short documents against longer ones. A new calculation of the similarity measure is proposed and some experimentation with the new measure is discussed.

## 1. Introduction

Information Retrieval Systems (Kowalski G., 1997) have been used for the retrieval of documents / information relevant to a submitted query. The queries are submitted either in Natural Language or in a Command Language. Modern Information Retrieval systems usually present the retrieved documents in decreasing order according to their similarity to the submitted query.

### 1.1 The measurement of similarity – The nearest neighbor problem

The problem of similarity for a document against a submitted query, also known as Document versus Query similarity or nearest neighbors has been the field of continuing research for more than 20 years (Smeaton A.F. & van Reijsbergen C.J., 1981). Many similarity functions / measures, and algorithms eliminating the query-document comparisons have been proposed (e.g. (Bentley J.L. et al, 1980), (Lucarella D., 1988), (Borlund & Ingweren, 1998)).

### 1.2 The vector space model

In the popular vector space model a data set of n unique terms is specified, called the index terms (or keywords) of the document collection, and every document can be represented by a vector,

$(T_1, T_2, \ldots, T_n)$

where $T_i$=1, if the index term i is present in the document, and 0 otherwise.

A query can be represented in the same manner. The document and query vectors can be envisioned as an n-dimensional vector space. A vector matching operation, based on the cosine correlation used to measure the cosine of the angle between

vectors can be used to compute the similarity. Hence, the following equation (e.g. (Lucarella D., 1988)) gives us a well-known method to measure the similarity of document $D_i$ against query Q:

$$S(D_i, Q) = \frac{\sum_{j=1}^{n} q_j t_{ij}}{\sqrt{\sum_{j=1}^{n} q_j^2 \cdot \sum_{j-1}^{n} t_{ij}^2}} = \frac{\sum_{j=1}^{n} q_j t_{ij}}{L_Q \cdot L_{D_i}} \quad (1)$$

where n is the number of index terms used in the collection, $t_{ij}$ is the weight of term j in document $D_i$ and $q_j$ is the weight of term j in the query.

The following two equations can be used to measure the terms $t_{ij}$ and $q_j$:

$$t_{ij} = 0.5 + 0.5 \cdot \frac{F_{ij}}{\max F_i} \quad (2)$$

$$q_j = \log_2 \left( \frac{N}{DOCFREQ_j} \right) \quad (3)$$

where $F_{ij}$ is the frequency of term j in document $D_i$, $maxF_i$ is the maximum frequency of the terms in document $D_i$, N is the number of documents in the collection and $DOCFREQ_j$ is the number of documents that include the index term j.

According to (1) some kind of the “document length” is given by:

$$L_{D_i} = \sqrt{\sum_{j-1}^{n} t_{ij}^2} \quad (4)$$

## 2. A new calculation of the similarity measure

The Computer Assisted Information Resources Navigation (CAIRN) system (Karanikolas N.N., 2007) is a general purpose Information Retrieval system based on the popular Vector Space Model (VSM) (Smeaton A.F. & van Reijsbergen C.J., 1981). CAIRN uses stems for automatic indexing of Greek, Latin and Multilingual texts. The queries are submitted in Natural Language and the retrieved documents are presented in decreasing order of similarity. Two nearest neighbor algorithms, based on Lucarella (Lucarella D., 1988), were implemented optimizing both the number of documents to be evaluated and the number of inverted lists to be inspected. The use of the system gave us an indication that the Nearest Neighbour Method (and the related similarity measure) used in CAIRN system presents a “preference” to short documents against longer ones. Hence, a new calculation of the similarity measure was proposed in a previous work related to medical data (Karanikolas N.N. & Skourlas C., 2000). To tackle the problem of "preference" of short documents against longer ones we decreased $L_D$ for longer documents suggesting that for the measurement of the similarity (see equation 1) is better to use equation (5) instead of the equation (4):

$$L_{D_i} = \ln(\sum_{j=1}^{n} t_{ij}^2 + e - 1) \quad (5)$$

As test base was selected a small collection (1033 medical documents, 30 queries, and the file of the relevance assessments) accessible through the IDOMENEUS technology transfer server at the University of Glasgow / Department of Computing Science (IDOMENEUS server). The experimentation with the two versions of the system confirmed our conjecture for the “preference” of the first calculation of the similarity measure to short documents against larger ones and that the proposed calculation of the measure improves the whole situation.

## 3. Results and discussion

### 3.1 An outline of the problem – Example

Let us consider the case of the following query (in Greek) submitted to the CAIRN system:

“Συγχώνευση Πλαστικών Μακεδονίας και Ομίλου Πετζετάκι”

The list of relevant documents retrieved by the CAIRN system version using the Lucarella measure contains a short document and a longer one. Short Document has four terms in common with the query, while longer document has five terms in common with the query. These documents follow:

*Short Document:*

> Συμφωνία υπέγραψε ο **όμιλος Πετζετάκι** και πιο συγκεκριμένα η θυγατρική του εταιρία **Πλαστικά Μακεδονίας** με τη Δημόσια Επιχείρηση Αερίου (ΔΕΠΑ) Α.Ε. Ειδικότερα, υπεγράφη σύμβαση για την προμήθεια της Δημόσιας Επιχείρησης Αερίου με σωλήνες πολυαιθυλενίου για το δίκτυο διανομής φυσικού αερίον, ύψους περίπου 1 δισ. δραχμές.Τη σύμβαση υπέγραψε ο διευθύνων σύμβουλος της "**Πλαστικά Μακεδονίας**" κ. Σταύρος Πλακαντωνάκτης παρουσία του υποδιευθυντού τεχνικών προμηθειών της ΔΕΠΑ κ. Στέφανου Οικονομίδη.

*Longer Document:*

> Μέχρι το τέλος της τρέχουσας χρήσης αναμένεται να ολοκληρωθεί η **συγχώνευση** της θυγατρικής εταιρίας **Πλαστικά Μακεδονίας**. του **ομίλου Πετζετάκι** με τη μητρική Α.Γ. **Πετζετάκις** ενώ παράλληλα θα προχωρήσει η αξιοποίηση της ακίνητης περιουσίας της θυγατρικής **Πλαστικά** Καβάλας. Τα παραπάνω ανέφερε ο κ.Γ. **Πετζετάκις**, πρόεδρος και διευθύνων σύμβουλος του ομίλου **Πετζετάκι**, στην Ένωση Θεσμικών Επενδυτών. Σε ό,τι αφορά τα οικονομικά αποτελέσματα της μητρικής εταιρίας, το 1996, τα κέρδη προφορών κινήθηκαν με ρυθμό αύξησης 41,09%, φθάνοντας τα 312,1 εκατ. δρχ. από 221,2 εκατ. δρχ. το 1995. Ο κύκλος εργασιών της εταιρίας εκτιμάται ότι θα ανέλθει σε 16,71 δισ. δρχ., παρουσιάζοντας αύξηση 26,2% από 13,24 δις. δρχ. το 1995. Οι συνθήκες ανταγωνισμού που επικρατούν στην αγορά έχουν οδηγήσει στη συμπίεση των περιθωρίων κέρδους και η κατάσταση αυτή αναμένεται να διαφοροποιηθεί μετά το β' εξάμηνο του 1997. Η διοίκηση εκτιμά ότι το 1997 τα κέρδη θα διπλασιαστούν και θα φθάσουν τα 640 εκατ. δρχ. ενώ ο κύκλος εργασιών θα προσεγγίσεί τα 20 δις. δραχμές. Ο **όμιλος Πετζετάκι** προχώρησε στη διετία 1995-1996 πρόγραμμα αναδιάρθρωσης με στόχο τη μείωση του κόστους παραγωγής και τη δημιουργία συνθηκών μακροχρόνιας ανάπτυξης. Το πρόγραμμα περιελάμβανε τη διακοπή των ζημιογόνων θυγατρικών (**Πλαστικά** Καβάλας), την υλοποίηση επενδυτικού προγράμματος με την εγκατάσταση του εργοστασίου των Αθηνών στη Θήβα, τον εκσυγχρονισμό του εργοστασίου των Θηβών και την έντονη δραστηριοποίηση στο κλάδο των πολυολεφινών. Η άλλη θυγατρική εταιρία του

**ομίλου Πετζετάκι**. που είναι εισηγμένη στο ΧΑΑ, τα **Πλαστικά Μακεδονίας**, παρουσίασαν το 1996 κύκλο εργασιών 6,4 δις. δρχ. από 5,7 δις. δρχ. το 1995 ενώ τα κέρδη προ φόρων μειώθηκαν σε 155 εκατ. δρχ. από 192 εκατ. δρχ. το 1995

The same query also submitted to the version of CAIRN with the new calculation of the similarity measure. The following table presents the relative position of these two documents according to both algorithms:

| Document | Position of Document according to the classic similarity measure | Position of Document according to the new version of similarity measure |
|---|---|---|
| Short | 1 | 2 |
| Long | 2 | 1 |

**Table 1:** Order of documents according to the classic similarity measure and our proposal

The following tables 2, 3 also present the comparative results for another simple query.

**Query: Είσοδος του ΟΤΕ στο χρηματιστήριο**

Entrance of Greek Telecommunication Organization (GTO) in Stock Exchanging

| | | | Repetitions of Query Terms in documents | | | |
|---|---|---|---|---|---|---|
| **Document number** | **Measurement of similarity using (5)** | **Fmax** | **Είσοδος (Entrance)** | **ΟΤΕ (GTO)** | **χρηματιστήριο (Stock Exchanging)** | **Document's physical size (lines)** |
| 4232 | 0.40 | 2 | 1 | 2 | - | 5 |
| 4917 | 0.36 | 4 | 1 | 3 | 4 | 18 |
| 502 | 0.33 | 7 | 1 | 4 | 4 | 42 |
| 4200 | 0.32 | 9 | 1 | 9 | 2 | 31 |
| 4355 | 0.31 | 3 | 2 | 1 | 1 | 18 |
| 4410 | 0.31 | 4 | 3 | 1 | 3 | 28 |
| 528 | 0.30 | 11 | 4 | 11 | - | 31 |
| 2543 | 0.27 | 5 | - | 5 | - | 7 |
| 3328 | 0.27 | 8 | - | 7 | 5 | 31 |
| 345 | 0.27 | 9 | - | 9 | 7 | 36 |
| 2481 | 0.26 | 11 | 3 | - | 11 | 52 |
| 4981 | 0.26 | 1 | 1 | - | 1 | 5 |
| 3465 | 0.26 | 2 | - | 1 | - | 3 |
| 2607 | 0.25 | 16 | 1 | - | 1 | 70 |
| 1874 | 0.25 | 8 | - | 8 | 1 | 45 |

**Table 2:** Order of documents according to our similarity measure proposal

| Query: Είσοδος του ΟΤΕ στο χρηματιστήριο<br>Entrance of Greek Telecommunication Organization (GTO) in Stock Exchanging | | | | | | |
|---|---|---|---|---|---|---|
| | | | Repetitions of Query Terms in documents | | | |
| **Document number** | **Measurement of similarity using (4)** | **Fmax** | **Είσοδος (Entrance)** | **ΟΤΕ (GTO)** | **χρηματιστήριο (Stock Exchanging)** | **Document's physical size (lines)** |
| 4232 | 0.28 | 2 | 1 | 2 | - | 5 |
| 4917 | 0.22 | 4 | 1 | 3 | 4 | 18 |
| 3465 | 0.21 | 2 | - | 1 | - | 3 |
| 2543 | 0.20 | 5 | - | 5 | - | 7 |
| 4355 | 0.18 | 3 | 2 | 1 | 1 | 18 |
| 502 | 0.18 | 7 | 1 | 4 | 4 | 42 |
| 4200 | 0.18 | 9 | 1 | 10 | 2 | 31 |
| 4228 | 0.18 | 3 | 1 | - | 2 | 2 |
| 528 | 0.18 | 11 | 4 | 11 | - | 31 |
| 4981 | 0.17 | 1 | 1 | - | 2 | 5 |
| 4410 | 0.17 | 4 | 3 | 1 | 3 | 28 |
| 3590 | 0.16 | 2 | 1 | - | 1 | 5 |
| 330 | 0.16 | 4 | - | 4 | - | 9 |
| 2459 | 0.16 | 4 | - | 3 | - | 7 |
| 3328 | 0.16 | 8 | - | 7 | 5 | 31 |

**Table 3:** Order of documents according the classic similarity measure

## 3.2 A detailed comparison

As we have already mentioned, the test bed for the proposed method is based on a collection of stock data. Table 4 illustrates our collection.

Three stockbrokers analyzed the results of their queries in the two different versions of the system and evaluated the results. Stockbrokers have evaluated the results of the second version (the one that uses equation 5) as better ones. The results of the submitted queries were also automatically saved for further study, in both versions of the system.

| $\sum_{j-1}^{n} t_{ij}^2$ | Number of documents |
|---|---|
| [1 - 10] | 287 |
| (10 – 20] | 901 |
| (20 – 30] | 1076 |
| (30 – 40] | 1234 |
| (40 – 50] | 806 |
| (50 – 60] | 346 |
| (60 – 70] | 156 |
| (70 – 80] | 79 |
| (80 – 90] | 30 |
| (90 – 100] | 24 |
| (100 – 110] | 19 |
| (110 – 120] | 8 |
| (120 – 130] | 3 |
| (130 – 140] | 6 |
| (140 – 150] | 3 |
| (150 – 200] | 3 |
| (200 - 286.667] | 3 |
| $\sum_{j-1}^{n} t_{ij}^2$ ranges from 1 to 286.677, its mean value is 33.462 and its variance is 370.452 | |

**Table 4:** The range of Σ in our stock exchange collection

The following table gives us a comparison of $L_D$ measured with equation (4) and equation (5):

| **Σ** | 1,00 | 2,00 | 3,00 | 4,00 | 5,00 | 6,00 | 7,00 | 8,00 | 9,00 | 10,00 | 15,00 | 20,00 | 30,00 | 40,00 | 50,00 | 70,00 | 90,00 | 100,00 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ln(Σ+e-1) (5)** | 1,00 | 1,31 | 1,55 | 1,74 | 1,90 | 2,04 | 2,17 | 2,27 | 2,37 | 2,46 | 2,82 | 3,08 | 3,46 | 3,73 | 3,95 | 4,27 | 4,52 | 4,62 |
| **sqrt(Σ) (4)** | 1,00 | 1,41 | 1,73 | 2,00 | 2,24 | 2,45 | 2,65 | 2,83 | 3,00 | 3,16 | 3,87 | 4,47 | 5,48 | 6,32 | 7,07 | 8,37 | 9,49 | 10,00 |

**Table 5:** $L_D$ measured with equation (4) and equation (5)

## 4. Conclusion and future work

In this paper, we suggest a new calculation of the measurement of a document's "length" in order to avoid the Nearest Neighbor method's preference for the short documents. We tested our method using a collection of documents with Σ varying (in most cases) from 1 to 100 and the results were positive.

CAIRN is a constantly evolving system. Currently we are working to conduct some more experimentation with the proposed, new similarity measure to confirm our

conjecture. A bigger Test Collection will be used, to give us a more accurate evaluation of the new calculation of the measure. We plan to evaluate the method using longer documents where Σ varies from 1 to 1000 or greater values. We have also some ideas in using other calculations of the similarity measure for improving the effectiveness of the system.

The modular design of the CAIRN information retrieval system has allowed us a great flexibility, and gives us the possibilities for future enhancements. Using modular design has also allowed us to implement, easily, different versions of the system and conduct experimentation with alternative similarity measures.

Hence, CAIRN system can be easily modified to incorporate other similarity measures and take advantage of user's suggested improvements.

## References


[1] Kowalski G: Information Retrieval Systems. Theory and Implementation, 1st edn (Printed in the USA; Kluwer Academic Publishers), ISBN 0-7923-9899-8, 1997.

[2] A.F. Smeaton and C.J. van Reijsbergen, 1981, The Nearest Neighbour problem in Information Retrieval. An algorithm using upperbounds. ACM SIGIR Forum 16, pp.83-87.

[3] Bentley, J. L., Weide, B.W. and Yao, A. C., 1980, Optimal expected time algorithms for closest point problems. ACM Trans. Math. Software, 6, pp. 563-580.

[4] Lucarella, D., 1988, A document retrieval system based on nearest neighbor searching. Journal of Information Science, 14, 25-33.

[5] Borlund and Ingweren, 1998, Measure of relative relevance and ranked half life: performance indicators for interactive IR. SIGIR'98, pp. 324-331.

[6] Karanikolas N.N. and Skourlas C., 2000, Computer assisted Information resources navigation. Medical Informatics and the Internet in Medicine, vol. 25, 133-146.

[7] Karanikolas N.N., 2007, Low cost, cross-language and cross-platform Information Retrieval and Documentation tools. Journal of Computing and Information Technology (CIT), ISSN: 1330-1136, volume 15, No 1.

[8] http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/
The med.all collection at the IDOMENEUS server.