HYPERBOLIC INFORMATION RETRIEVAL

Júlia Góth
Department of Computer Science, University of Veszprém,
Egyetem u. 10, 8200 Veszprém, Hungary

Abstract

There is one Information Retrieval model that uses geometrical space: it is the Vector Space Model, which is defined in Euclidean Space. The paper shows that it is possible to define a Vector Space Model in non-Euclidean Space, too. Namely, the paper proposes a Vector Space Model over the Cayley-Klein Hyperbolic Geometry using a similarity measure derived from the hyperbolic distance. It is shown that the proposed model is equivalent with the classical Vector Space Model using a normalized weighting scheme. Experiments are also reported to demonstrate the model suggested.

Key words

Information Search and Retrieval, Euclidean Geometry, Hyperbolic Geometry, Cayley-Klein Model, Similarity measures, Ranking order preservation

1. INTRODUCTION

In the field of Information Retrieval (IR), the Vector Space Model (VSM) is an important, well-understood and extensively researched classical model, which has been widely used to process texts efficiently and retrieve information for some forty years (Salton, 1966). The VSM is called so because each document and query is mapped to a point in the feature space based on frequencies of keywords appearing in the text. The feature space is mathematically modelled by the orthonormal Euclidean space, i.e., the space (or geometry) defined by a system of pairwise perpendicular coordinate axes.

So far, the Euclidean geometry is the only type of space used in the VSM in general, but non-Euclidean Geometry is becoming increasingly important in modern science and technology. Thus, one may raise the following question: Can non-Euclidean spaces — which are mathematically consistent alternatives to Euclidean geometry, and have

been developed about one and a half century ago and more than half a century before the vector space model — be used to model the VSM, and thus be applied to IR?

The application of non-Euclidean spaces to information processing in general seems to experience its beginnings: they are used for information visualisation. In a non-Euclidean space the area of a circle grows exponentially with respect to its radius, whereas in Euclidean space the area only grows quadratically. Thanks to this property a convenient way to visualize exponentially growing trees can be derived (Phillips and Gunn, 1992; Phillips, Levy and Munzner, 1993). They draw 3D hyperbolic pictures of large hierarchies or graphs (such as the Web) in the interior of a ball, use Euclidean straight lines, but the way distance is measured is changed. Thus, an effective way to visualise structures is obtained (more can be represented in less space, although in a distorted way; in a fisheye view style).

The present paper addresses the question above, and shows that the answer is yes. It investigates a possibility to apply non-Euclidean spaces to IR by defining a VSM in the hyperbolic space with a hyperbolic similarity measure. Experimental results are also reported, and it is shown that the new model is equivalent with the Cosine-based VSM with normalised weighting scheme.

2. VECTOR SPACE MODEL AND σ-SPACE

2. 1 Classical Vector Space Model

Given *documents* $D_j, j = 1,...,m \in \aleph$, and *terms* t_i, $i = 1,...,n \in \aleph$. In the *Vector Space Model* (van Rijsbergen, 1979; Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999) of Information Retrieval, every document D_j is assigned a vector $\mathbf{w}_j = (w_{ij})_{i=1,...,n}$ of *weights*, where $w_{ij} \in \Re$ denotes the *weight* of term t_i for document D_j. The matrix $W = (w_{ij})_{n \times m}$ is called the *term-by-document matrix*. The general form of the *weighting scheme* (Berry and Browne, 2000) is as follows:

$$w_{ij} = \text{local\_weight}_{ij} \times \text{global\_weight}_i \times \text{normalisation}_j = l_{ij} \times g_i \times n_j$$

Let Q denote a *query* coming form the user, and $\mathbf{q} = (q_i)_{i=1,...,n}$ the corresponding query vector. The vectors $\mathbf{w}_j$ and $\mathbf{q}$ belong to the E_n Euclidean orthonormal space, in which the weights $\mathbf{w}_j$ and $\mathbf{q}$ are regarded as Cartesian coordinates (of points corresponding to D_j and Q).

2.2 Similarity measures

The *relevance* of document D_j relative to Q is given by the value of a *similarity measure* $\sigma(\mathbf{w}_j, \mathbf{q})$, whose general form is as follows:

$$\sigma = \frac{\mathbf{w}_j\mathbf{q}}{\Delta}$$

where $\mathbf{w}_j\mathbf{q}$ denotes the inner product of the vectors $\mathbf{w}_j$ and $\mathbf{q}$.

Depending on the formula of Δ, the similarity measures proposed are as follows (Dominich, 2001):

Dot product: $\Delta = 1$

Cosine measure: $\Delta = ||\mathbf{w}_j|| \cdot ||\mathbf{q}||$

Dice's coefficient: $\Delta = \mathbf{w}_j + \mathbf{q}$

Jaccard's coefficient: $\Delta = \sum_{i=1}^{n} \frac{w_{ij} + q_i}{2^{w_{ij} q_i}}$

Overlap coefficient: $\Delta = \min(\sum_{i=1}^{n} w_{ij}, \sum_{i=1}^{n} q_i)$

Because, in general the similarity measures do not preserve the rank order of the retrieved documents, perhaps only practice and experimentation, but no sound theoretical argument can recommend which one to use to obtain better results. In practice the most commonly used or proposed similarity measure is the Cosine measure.

2.3 Similarity space (σ-space)

The underlying common and formal properties of similarity measures are as follows:

1. Symmetry. The order in which the query and the document are considered when computing the similarity value is indifferent. Formally:

   $\sigma(a, b) = \sigma(b, a), \forall a, b$ (symmetry)

2. Reflexivity. The value of the similarity measure is equal to a predefined and fixed maximal value κ if the query and the document are exactly the same. (The reverse is not necessarily true.) For example, if σ is normalised then κ may be taken as being equal to 1. Formally:

   $a = b \Rightarrow \sigma(a, b) = \kappa$ (reflexivity)

The concept of a σ-space, introduced in (Dominich, 2001), is a formal generalisation of the VSM in order to emphasise the fact that retrieval is based on similarity measures. A set D of objects with a symmetric and reflexive similarity measure, i.e.,

$\sigma: D \times D \rightarrow \mathfrak{R}$

$\sigma(a, b) = \sigma(b, a), \forall a, b \in D$

$a = b \Rightarrow \sigma(a, b) = \kappa$, (where κ is a predefined and fixed maximal value)

is called a σ-*space*. It can be shown that the following theorem holds:

THEOREM
Let $\langle E, \mu \rangle$ be a (pseudo-) metric space (μ is normalised, which is always possible). Then

1. The induced topological space is a σ-space on E.
2. $\langle E, 1 - \mu \rangle$ is a σ-space. ♦

Because the different similarity measures do not preserve the rank order, and taking into account the theorem above, which allows for deriving a similarity measure from a (pseudo) metric, the quest for exploring similarities from metrics in spaces (geometries) other than Euclidean is addressed.

3. CAYLEY-KLEIN HYPERBOLIC GEOMETRY

Non-Euclidean Geometry is different form Euclidean Geometry in that, the axiom of Parallels (in plane, there exists exactly one parallel line to a given line through a given point that is not on the given line) is not valid. One possible non-Euclidean Geometry is Cayley-Klein hyperbolic space (C-KHS) or model (Bolyai, 1987), where: $\mathfrak{R}^n$ (Császár, 1974) denote the Euclidean (orthonormal) space. Let A and B denote two points in $\mathfrak{R}^n$, and let $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ denote their Cartesian coordinates, respectively. The Euclidean distance $d_E(AB)$ between the points A and B is as follows (Patterson and Rutherford, 1965):

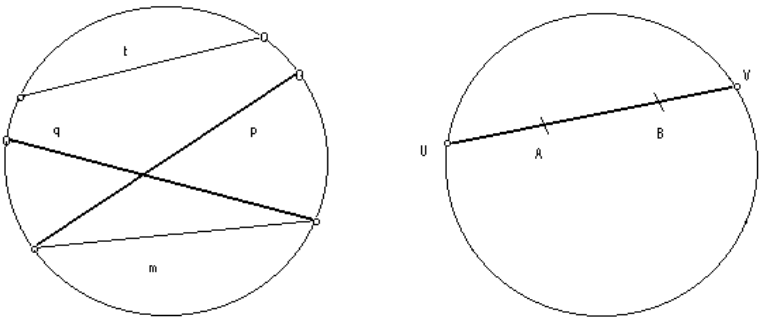
$$d_E(AB) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Let

$$S(O, r) = \{P(x_1, x_2, ..., x_n) \mid \sum_{i=1}^{n} x_i^2 < r^2 \}$$

denote the interior of a hyper-sphere S having its centre in the origin O of space $\mathfrak{R}^n$, and radius $r \in \mathfrak{R}$, $r > 0$.

The *space* C-KHS is $S(O, r)$. The *points* P of the C-KHS are all the points of $S(O, r)$, i.e, $P \in S(O, r)$. The *lines* of the C-KHS space are open chords of the hyper-sphere S deprived of its endpoints. If the lines m and n have one common endpoint (on the boundary of the hyper-sphere in the Euclidean space) they are referred to as *asymptotically parallel* (Figure 1.a). The scientific role and importance of non-Euclidean geometries is well-known (Anderson, 1999). In our context the hyperbolic distance rather than parallelism will play an important role.



a) b)

Figure - 1.

a) m, p, q, t are lines in the C-KHS. Notice that the endpoints do not belong to the C-KHS, nor does the points of the circle (the circle is only drawn to show the 'limits' of the hyperbolic space). The lines p and q are asymptotically parallel to line m: m || p, m || q. The line t is divergently parallel to line m: m || t.

b) Example for a line segment AB in the C-KHS.

The C-KHS space satisfies Hilbert's axioms on incidence, ordering and congruence, as well as Archimedes' and Cantor's axioms on continuity (Hilbert and Cohn-Vossen, 1932), and is thus a continuous absolute space. The concept of distance is defined using that of cross ratio The hyperbolic length $d_H(AB)$ of the line segment AB is defined as the cross-ratio of the points U, A, B, V:

$$d_H(AB) = k \cdot \left| \ln \frac{d_E(AU) \cdot d_E(BV)}{d_E(AV) \cdot d_E(BU)} \right| \qquad (1)$$

where $k \in \mathfrak{R}$ is a positive constant, U and V are the points of intersection of the Euclidean line through the points A and B with the hyper-sphere (Figure 1b). In the following we will assume, without loss of generality, that k = 1. The following is a remainder of those well-known properties,

which are of interest for us. The hyperbolic distance is non-negative (immediate from equation given above), i.e.,

$$d_H(AB) \geq 0, \forall\ A, B \in \text{C-KHS} \tag{2}$$

The hyperbolic distance is symmetric:

$$d_H(AB) = \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} \right| = \left| \ln \frac{1}{\frac{d_E(AV)\cdot d_E(BU)}{d_E(AU)\cdot d_E(BV)}} \right|$$

$$= \left| \ln 1 - \ln \frac{d_E(AV)\cdot d_E(BU)}{d_E(AU)\cdot d_E(BV)} \right| = \left| \ln \frac{d_E(AV)\cdot d_E(BU)}{d_E(AU)\cdot d_E(BV)} \right|$$

$$= d_H(BA), \quad \forall\ A, B \in \text{C-KHS} \tag{3}$$

The hyperbolic distance is reflexive:

$$A = B \Rightarrow d_H(AB) = d_H(AA) \Rightarrow$$

$$\Rightarrow d_E(BU) = d_E(AU), \text{ and } d_E(BV) = d_E(AV) \Rightarrow |\ln 1| = 0 \tag{4}$$

The hyperbolic distance satisfies the triangle inequality:

$$d_H(AB)+d_H(BC) = \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} \right| + \left| \ln \frac{d_E(BU)\cdot d_E(CV)}{d_E(BV)\cdot d_E(CU)} \right| \geq$$

$$\geq \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} + \ln \frac{d_E(BU)\cdot d_E(CV)}{d_E(BV)\cdot d_E(CU)} \right| =$$

$$= \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} \times \frac{d_E(BU)\cdot d_E(CV)}{d_E(BV)\cdot d_E(CU)} \right| =$$

$$= \left| \ln \frac{d_E(AU)\cdot d_E(CV)}{d_E(AV)\cdot d_E(CU)} \right| = d_H(AC) \qquad \forall A,B,C \in \text{C-KHS} \tag{5}$$

Due to these properties the hyperbolic distance is a metric. Additional properties of the hyperbolic distance are as follows:

(i) The hyperbolic and the Euclidean distances preserve the ranking order, if A is the origin of hyper-sphere S, i.e.,

Assume that,

$$d_E(AB) < d_E(AC), \Rightarrow \frac{r - d_E(AB)}{r + d_E(AB)} > \frac{r - d_E(AC)}{r + d_E(AC)} \Rightarrow$$

$$\Rightarrow \left| \ln \frac{r - d_E(AB)}{r + d_E(AB)} \right| < \left| \ln \frac{r - d_E(AC)}{r + d_E(AC)} \right| \quad \text{accordingly,}$$

$d_H(AB) < d_H(AC)$, because,

$$d_H(AB) = \left| \ln \frac{d_E(AU) \cdot d_E(BV)}{d_E(AV) \cdot d_E(BU)} \right| = \left| \ln \frac{r \cdot (r - d_E(AB))}{r \cdot (r + d_E(AB))} \right| \tag{6}$$

(ii) d_H becomes infinitely large when either of the points approaches the surface of the hyper-sphere, i.e.,

$$B \to V \Rightarrow d_E(BV) = 0 \Rightarrow \lim_{B \to V} d_H(AB) = +\infty \tag{7}$$

4. HYPERBOLIC INFORMATION RETRIEVAL MODEL

4.1 Hyperbolic Space of Documents

Given a VSM. Let $\mathfrak{R}'^n$ denote the n-dimensional Euclidean space obtained by translating the space $\mathfrak{R}^n$ into Q, i.e., the origin O of $\mathfrak{R}^n$ is translated into Q. Let us consider the following C-KHS:

$$S'(O' = Q(q_1, q_2, ..., q_j, \ldots, q_m), r), \quad r > \max_D d_E(QD)$$

The hyperbolic distance $d_H(QD)$ in $S'(Q, r)$ is as follows:

$$d_H(QD) = \left| \ln \frac{d_E(QU) \cdot d_E(DV)}{d_E(QV) \cdot d_E(DU)} \right| = \left| \ln \frac{r \cdot (r - d_E(QD))}{r \cdot (r + d_E(QD))} \right| =$$

$$= \left| \ln \frac{r - \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}}{r + \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}} \right| \qquad (8)$$

The classical similarity measures in the VSM never take on negative values, they only have positive values, and from a psychological point of view, this seems normal (a likeness in meaning between objects should be expressed by a positive quantity). From this point a modified hyperbolic distance: δ_H ($\delta_H = d_H / (1 + d_H)$) has been introduced.

Because the hyperbolic distance d_H is a metric, the C-KHS space $S'(Q, r)$ can be turned into a σ-space (Theorem) by defining the following similarity measure $\sigma_H(\mathbf{w}, \mathbf{q})$:

$$\sigma_H(\mathbf{w}, \mathbf{q}) = 1 - \delta_H(QD),$$

Thus, the explicit form of the hyperbolic similarity measure $\sigma_H(\mathbf{w}, \mathbf{q})$ is as follows:

$$\sigma_H(\mathbf{w},\mathbf{q}) = \frac{1}{1 + \left| \ln \frac{r - \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}}{r + \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}} \right|} =$$

$$= \frac{1}{\ln \frac{e}{\frac{r - \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}}{r + \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}}}} =$$

$$= \left(\ln \left(e \cdot \frac{r + \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}}{r - \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}} \right) \right)^{-1} \qquad (9)$$

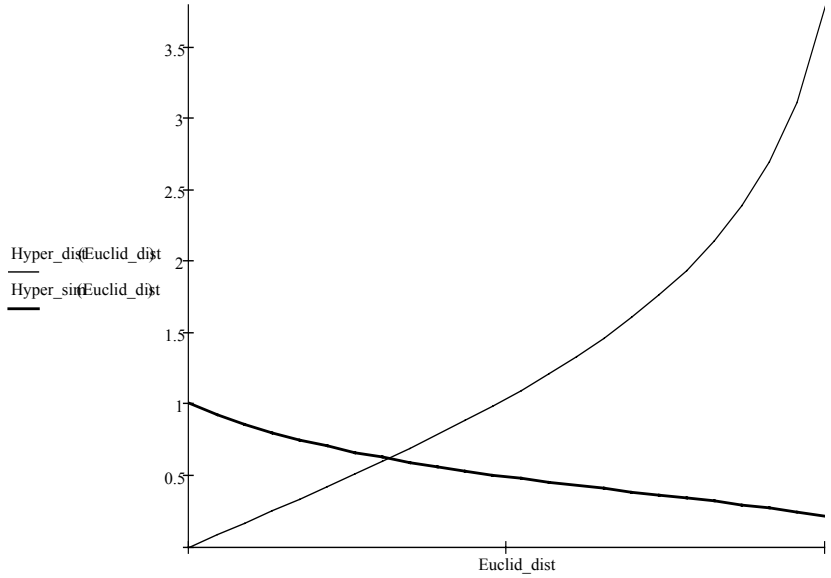


Figure -2. Plots of hyperbolic distance and similarity as functions of Euclidean distance

The function $\sigma_H(\mathbf{w}, \mathbf{q})$ is a similarity because it satisfies the three similarity properties (Van Rijsbergen, 1979), i.e.,

(i) Normalization. $0 \le \sigma(w, q) \le 1$, because (Figure 2.):

$$0 \le \left(\ln\left(e \cdot \frac{r + \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}}{r - \sqrt{\sum_{j=1}^{m} (w_{ji} - q_j)^2}} \right) \right)^{-1} \le 1 \qquad (10)$$

(ii) Symmetry. The order in which the query and the document are considered when computing the similarity value is indifferent. Formally: $\sigma(a, b) = \sigma(b, a)$, $\forall a,b \in D$, because:

$$d_E(QD) = d_E(DQ) \Rightarrow \sigma_H(Q, D) = \sigma_H(D, Q) \qquad (11)$$

(iii) Reflexivity. The value of the similarity measure is equal to a predefined and fixed maximal value κ if the query and the document are exactly the same. For example, if σ is

normalised then κ may be taken as being equal to 1. Formally:

$a = b \Rightarrow \sigma(a, b) = \kappa$ (reflexivity) $\forall a,b \in D$, because:

$Q = D \Rightarrow d_E(QD) = d_E(QQ) = 0 \Rightarrow$

$$\Rightarrow \sigma_H(Q, Q) = (\ln(e \cdot \frac{r}{r}))^{-1} = (\ln e)^{-1} = 1 \qquad (12)$$

4.2 Additional properties of the hyperbolic similarity measure $\sigma_H(w, q)$

(i) From the projective aspects, it can easily be proved, that enlarging the radius of the hyper-sphere, the documents seem to be similar, because we could not distinguish them, i.e.,

$$\lim_{r\to\infty} \sigma_H(w,q) = (\ln(e))^{-1} = \ln e = 1 \qquad (13)$$

Consequently, in the hyperbolic space, the radius should not be too large, but larger anyway than the $\max_D d_E(QD)$. So, the radius:

$r := \max_D d_E(QD) + \varepsilon$, where ε is optional, and $0 < \varepsilon << +\infty$

(ii) According to the property (7) the hyperbolic similarity measure becomes zero when either of the points approaches the surface of the hyper-sphere, i.e.,

$$B \to V \Rightarrow \lim_{B\to V} d_H(AB) = +\infty \Rightarrow$$

$$\lim_{B\to V} \sigma_H(AB) = \lim_{B\to V} (\ln(e \cdot \frac{r + d_E(AB)}{r - d_E(AB)}))^{-1} = \lim_{B\to V} 1 - \frac{d_H(AB)}{1 + d_H(AB)} = 0 \qquad (14)$$

5. EXPERIMENTS

The hyperbolic information retrieval model (HM) was tested on a medical database as regards rank order preservation, to illustrate the formally proof of it.

5.1 Rank order preservation of the Cosine and Hyperbolic Similarity Measures

Let $d_{i,j}$ mean, for example, the number of occurrences of term t_i in document D_j. (d_{ij} could be any other appropriate value, in fact.) The normalised weighting scheme means that the terms are assigned normalised weights w_{ij} as follows:

$$w_{ij} = \frac{d_{ij}}{\sqrt{\sum_{i=1}^{n} d_{ij}^2}}$$

The query terms are assigned weights similarly. Under this weighting scheme, the Cosine and Hyperbolic measures become:

$$Cosine_j = \sum_{i=1}^{n} w_{ij} q_i \quad \text{and} \quad \sigma_j = \left(\ln\left(e \cdot \frac{r + \sqrt{2 - 2\sum_{i=1}^{n} w_{ij} q_i}}{r - \sqrt{2 - 2\sum_{i=1}^{n} w_{ij} q_i}} \right) \right)^{-1}$$

respectively, because $\sum_{i=1}^{n} w_{ij}^2 = 1, \sum_{i=1}^{n} q_i^2 = 1$. We have:

$$\sum_{i=1}^{n} w_{ij} q_i \le \sum_{i=1}^{n} w_{ik} q_i \Leftrightarrow \sqrt{2 - 2\sum_{i=1}^{n} w_{ij} q_i} \ge \sqrt{2 - 2\sum_{i=1}^{n} w_{ik} q_i}$$

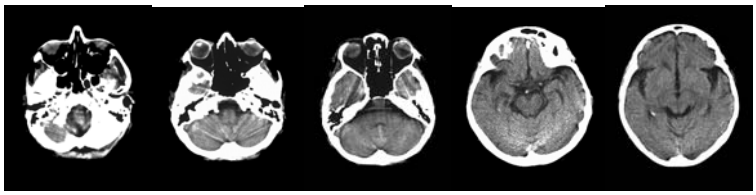
and thus $Cosine_j \le Cosine_k \Leftrightarrow \sigma_j \le \sigma_k$.

This means that the Cosine and Hyperbolic measures preserve the rank order of the documents. The following part reports on application and experiment illustrating rank order preservation.

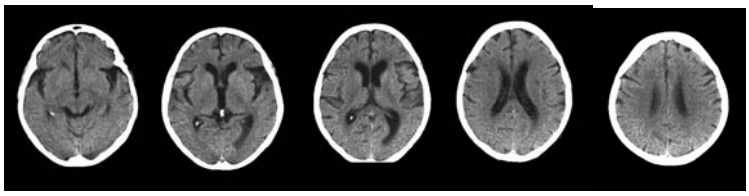
5.2 The Medical Database

The database contained 40 cases including two parts: (i) CT images (each case contains from 10 to 14 image slices), (ii) textual information: scanning time, patient age, patient sex, patient notes, paresis information, and case report (the demographic data is not necessarily real to ensure anonymity). Figure 3. shows an example.

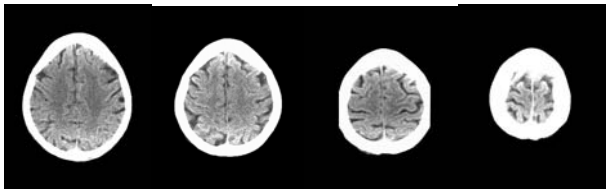
Slices (1 through 14: cross sections from the bottom to the top of the head)



1 2 3 4 5



6 7 8 9 10



11 12 13 14

Textual Information
Patient: 70-year old woman

Patient notes: She was not aphasic, and was fully conscious. The patient had a severe left-sided hemiparesis.

Case Riport: There are no signs of hyperdensity, large infarct or hyperdense artery. The infarct extent is under 33%. Patient suitable for thrombolysis.

Figure - 3. A patient's case in the database (example for illustration purposes only): fourteen slices and a short radiological report.

The cases were indexed using relevant medical terms in the written reports and a set of criteria relative to image content (Figure 4).

Aphasia sudden global Bedridden Cardiac arhytmia Collapse sudden Coma Consciousness undisturbed impaired Coronary artery stenosis	Eye deviation conjugated Palsy gaze Myocardial infarction Orientation undisturbed impaired completely disturbed Somnolence Stupor	Hemiparesis progressive severe sudden slight moderate very severe global facial subacute left-sided right-sided

a) Examples of relevant medical terms in written reports used as index terms. For example, the term 'palsy gaze' has Boolean values (Yes, No), whilst the term 'moderate hemiparesis' has the weight 0.5 in the original document-term matrix.

Hyperdensity, Haemorrhage, Infarction, Hyperdense artery, Hypodensity, Thrombolysis, Tissue volume, Deformation, Brain swelling

b) Criteria expressing relevant image features.

Figure -4. Examples of relevant medical terms used as index terms and criteria expressing image features.

5.3 Rank order preservation

Both the medical terms and criteria were assigned numerical values (see Figure 3 for examples). Thus, a document-term matrix D was constructed, where $D_{i,j}$ denoted the numeric value assigned to term (or criteria) i for case j (corresponding to a 'document' D_j). As suggested in (Meadow, Boyce and Kraft, 1999; Berry and Browne, 2000) for technical disciplines (like human brain CT imaging), the *tfn* (normalized weighting) scheme was used and tested (m denotes the total number of terms and criteria, n denotes the total number of cases). The aim of the experiment was to compare the rankings of the Hyperbolic and Cosine similarity measures, which is the mostly used and recommended. Figure 5. shows the results were obtained, which confirmed what had been expected based on the theoretical considerations above.

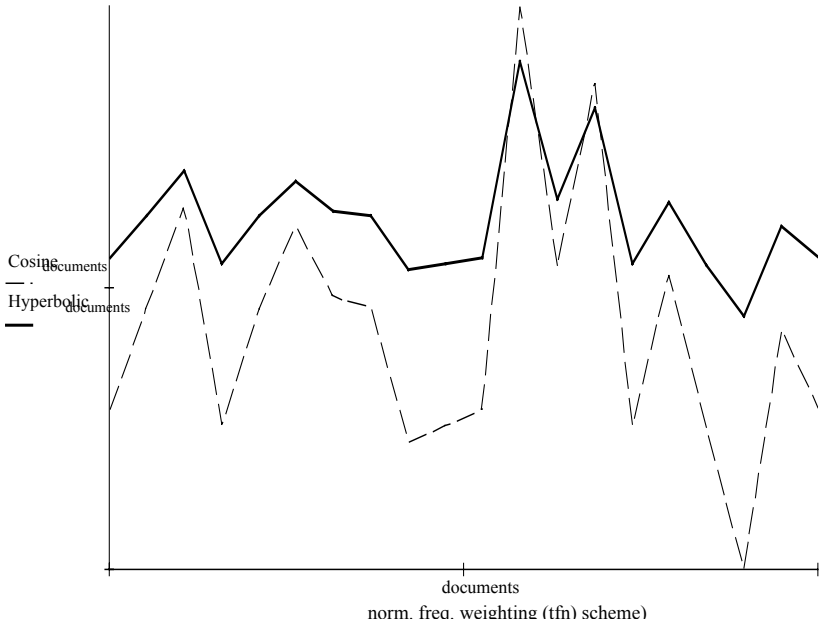


Figure -5. Comparison of similarity measures. The Cosine and Hyperbolic measures preserve the rank order under the normalised weighting scheme.

6. CONCLUSIONS AND FUTURE WORKS

The Euclidean geometry is the only type of space used in the VSM. The paper proposed that it is possible to define a Vector Space Model over the Cayley-Klein Hyperbolic Geometry. After presenting the classical Vector Space Model, and the Cayley-Klein Geometry with their properties, the Hyperbolic Information Retrieval model was defined using a similarity measure, derived from the hyperbolic distance. It was shown that this model was equivalent with the traditional Vector Space Model using a normalized weighting scheme, because both preserved the ranking order of the retrieved documents. Experiments were also reported implementing the proposed hyperbolic VSM.

In my paper it was shown that the Vector Space Model could be extended to the non-Euclidean Space thus far without any additional property for the Information Retrieval. Since the non-Euclidean geometry has important role in modern science and technology, especially in the visualization, my future plan is to investigate the effects of the Hyperbolic Model in the field of „Information Retrieval in context”.

7. ACKNOWLEDGEMENT

This research was sponsored by the National Research and Development Fund of the Széchenyi Plan, Hungary, Consortium NKFP OM #2/052/2001, and also by grants AKP #2001-140, and OTKA #T030194. Views and conclusions contained in this paper are those of the author, and should not be interpreted as necessarily representing those of the Consortium above.

The author would like to thank physician radiologist Dr. P. Barsi for providing the brain CT database, and M.Sc. students T. Kiezer and R. Szlávik for performing the experiments.

8. REFERENCES

[1] Anderson, J.W. (1999). *Hyperbolic Geometry*. Springer Verlag, New York.

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

[3] Berry, M.W. and Browne, M. (2000). *Understanding Search Engines – Mathematical Modeling and Text Retrieval*. SIAM, Philadelphia.

[4] Bolyai, J. (1987). *APPENDIX: The Theory of Space*. Akadémiai Kiadó, Budapest (Eds.: Kárteszi, F. and Szénássy, B.)

[5] Császár, Á. (1974). *General Topology*. Akadémiai Kiadó, Budapest.

[6] Dominich, S. (2001). *Mathematical Foundations of Information Retrieval*. Kluwer Academic Publishers, Dordrecht, Boston, London.

[7] Hilbert, D. and Cohn-Vossen, S. (1932). *Anschauliche Geometrie*. Springer Verlag, Berlin-Heidelberg-New York.

[8] Mark Phillips and Charlie Gunn (1992). Visualizing hyperbolic space: Unusual uses of 4x4 matrices. In *1992 Symposium on Interactive 3D Graphics (Boston, MA, March 29 - April 1 1992)*, volume 25, pages 209-214, New York. ACM SIGGRAPH. special issue of *Computer Graphics*.

[9] Mark Phillips, Silvio Levy, and Tamara Munzner (1993). Geomview: An interactive geometry viewer. *Notices of the American Mathematical Society*, 40(8):985-988, October 1993. Computers and Mathematics Column.

[10] Meadow, C.T., Boyce, B.R. and Kraft, D.H. (1999). *Text Information Retrieval Systems*. Academic Press, San Diego, San Francisco, New York, Boston, London, Sydney, Tokio.

[11] Petterson, E.M. and Rutherford, D.E. (1965). *Einführung in die Abstracte Algebra*. Bibliographisches Institut, Mannheim.

[12] Salton, G. (1966). Automatic Phrase Matching. In Hayes, D.G. (ed.) *Readings in Automatic Language Processing*. American Elsevier Publishing Company, Inc., New York, pp. 169-188.

[13] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill, New York.

[14] Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois.

[15] Van Rijsbergen, C.J. (1979). *Information Retrieval*. Butterworth, London.