Application for participation at the Doctoral Forum

Last name: Zhou

First name: Qiuju

Institutional affiliation (university/department/link to website):
National Science Library, Chinese Academy of Sciences http://www.las.ac.cn/
Full address of the doctoral student including phone and fax numbers and email address:
33 Beisihuan Xilu, Zhongguancun , Beijing P.R.China ,100190.Telphone：8601082627304，Fax：8601082627304，zhouqj@mail.las.ac.cn
Names of the supervisor(s) (links to their websites):
Fuhai Leng http://iauthor.las.ac.cn/CN/0000-0001-6306-1685

(in total max 1500 words)
Description of doctoral research project (including research questions, theoretical background, planned methodology, current status)

1 Research Questions

I was inspired through direct communication with Loet Leydesdorff and Ismael Rafols, find that Rao-Stirling diversity and the similarity-weighted Cosine cannot follow the rules of Euclidean space, but does follow the rules of Affine Space. I am trying to discuss the nature of the various diversity and similarity measures from the perspective of Analytic Geometry.

2 Theoretical Framework

The Basic Geometric Representation in Affine Space

Two matrices are used to calculate diversity and similarity in affine space (Fig. 1). The first matrix (occurrence matrix) describes the occurrence of Category Entities (columns) in systems (rows). The second contains similarities or dissimilarities between Category Entities.

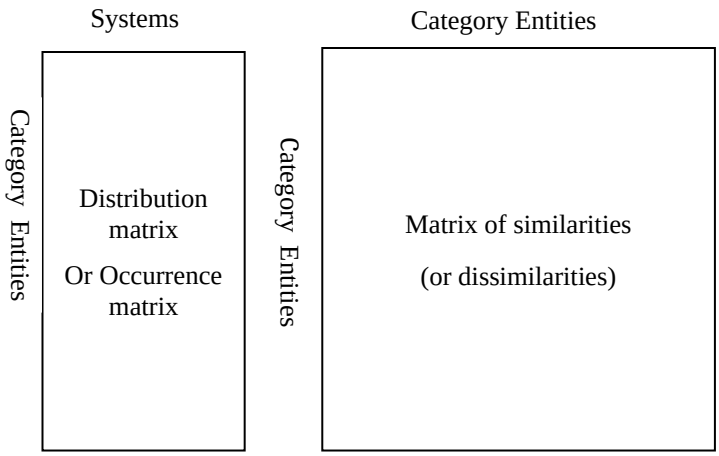

Figure 1 the basic data: Distribution or Occurrence matrix and matrix of similarities/dissimilarities

Category Entities are represented as axes in n-dimensional affine space *A*. If the number of Category Entities is n, there are n axes, named $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n$. The similarity matrix of Category Entities composes an n-dimensional *affine coordinate system*. The angle of each pair of axes is not always orthogonal. An *affine coordinate system* $B \equiv (O; \vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)$ consists of an origin *O* in *A* and a basis $\vec{a}_i (i = 1, \ldots, n)$ of the underlying vector space V.

Geometrically, the systems can be represented by points in space, using the distribution in Category Entities of a system (in the columns of the matrix) as coordinates. We can also graphically represent a vector by simply drawing an arrow from the origin *O* to the point in the space.

The Rules of Vector Addition and Inner Product within Affine Coordinate Systems

In Two-dimensional Affine Plane

If *A* is a two-dimensional space, two non-collinear bases $\vec{a}_1$, $\vec{a}_2$, representing two Category Entities, constitute the plane of the axis $\vec{a}_1$, axis $\vec{a}_2$. This coordinate system is called affine plane coordinate system (the angle of $\vec{a}_1$ and $\vec{a}_2$ is θ). Setting the coordinate origin to *O*, if a point X (representing a system) within the coordinate system satisfies $\overrightarrow{OX} = \vec{X} = x_1\vec{a}_1 + x_2\vec{a}_2$, x_1, x_2 are the length of $\vec{X}$ in the direction of basis $\vec{a}_1, \vec{a}_2$, the predetermined coordinates of the point X (x_1, x_2) within this coordinate has the following operational rules established:
The vectors $\vec{X} = (x_1\vec{a}_1, x_2\vec{a}_2), \vec{Y} = (y_1\vec{a}_1 + y_2\vec{a}_2)$ are shown geometrically in Figure 2.

Hence, Cartesian coordinates are a very special kind of affine coordinates that correspond to the case where $\vec{a}_1$ =(1,0), $\vec{a}_2$ =(0,1).

$\vec{X} = x_1\vec{a}_1 + x_2\vec{a}_2$, $\vec{Y} = y_1\vec{a}_1 + y_2\vec{a}_2$

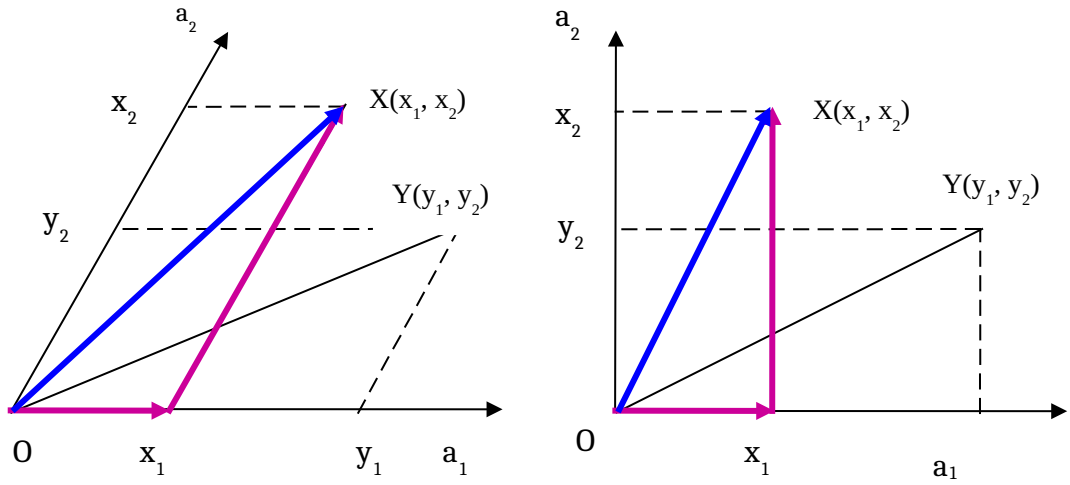

Figure 2 Two-dimensional affine coordinates contains Oblique coordinates and Cartesian (Orthogonal) coordinates

The inner product of vectors $\vec{X}$ and $\vec{X}$ should be:

$$\vec{X}\cdot\vec{Y} = (x_1\vec{a}_1 + x_2\vec{a}_2)\cdot(y_1\vec{a}_1 + y_2\vec{a}_2) = x_1y_1\vec{a}_1\cdot\vec{a}_1 + x_1y_2\vec{a}_1\cdot\vec{a}_2 + x_2y_1\vec{a}_2\cdot\vec{a}_1 + x_2y_2\vec{a}_2\cdot\vec{a}_2$$

(1)

$$\vec{X}\cdot\vec{X} = |\vec{X}|^2 = (x_1\vec{a}_1 + x_2\vec{a}_2)\cdot(x_1\vec{a}_1 + x_2\vec{a}_2) = x_1x_1\vec{a}_1\cdot\vec{a}_1 + x_1x_2\vec{a}_1\cdot\vec{a}_2 + x_2x_1\vec{a}_2\cdot\vec{a}_1 + x_2x_2\vec{a}_2\cdot\vec{a}_2$$

(2)

$$\vec{Y}\cdot\vec{Y} = |\vec{Y}|^2 = (y_1\vec{a}_1 + y_2\vec{a}_2)\cdot(y_1\vec{a}_1 + y_2\vec{a}_2) = y_1y_1\vec{a}_1\cdot\vec{a}_1 + y_1y_2\vec{a}_1\cdot\vec{a}_2 + y_2y_1\vec{a}_2\cdot\vec{a}_1 + y_2y_2\vec{a}_2\cdot\vec{a}_2$$

(3)

And where $|\vec{X}|^2$ and $|\vec{Y}|^2$ are the squares of the norms of $\vec{X}$ and $\vec{Y}$.

The included angle of the axis $\vec{a}_1$ and $\vec{a}_2$ is cos $(\vec{a}_1,\vec{a}_2)$. The vector parallel with the axis $\vec{a}_1$ or $\vec{a}_2$ has the same direction with $\vec{a}_1$ or $\vec{a}_2$. So,

$\vec{a}_1 \cdot \vec{a}_1=\cos(\vec{a}_1,\vec{a}_1)=1$, $\vec{a}_2 \cdot \vec{a}_2=\cos(\vec{a}_2,\vec{a}_2)=1$, $\vec{a}_1 \cdot \vec{a}_2=\cos(\vec{a}_1,\vec{a}_2)$

Change the formula $\cos(\vec{a}_i,\vec{a}_j)=\frac{\vec{a}_i \cdot \vec{a}_j}{|a_i||a_j|}$, we get, $\vec{a}_i \cdot \vec{a}_j = |a_i||a_j|\cos(\vec{a}_i,\vec{a}_j)$

a_i, a_j are the bases, so the length of the basis is 1.The norm of basis $|a_i| = |a_j| = 1$. Then,

$a_i \cdot a_j = \cos(a_i,a_j)$, $a_i^2 = \cos(a_i,a_i) = 1$, $a_j^2 = \cos(a_j,a_j) = 1$

The Inner Product of vectors $\vec{X}$ and $\vec{X}$ can be written in the following form:

$$\vec{X} \cdot \vec{Y} = \sum_{i,j=1}^{2} x_i y_j \cos(\vec{a}_1,\vec{a}_2) = \sum_{i,j=1}^{2} x_i y_j S_{ij} \quad (4)$$

$$\vec{X} \cdot \vec{X} = \sum_{i,j=1}^{2} x_i x_j \cos(\vec{a}_1,\vec{a}_2) = \sum_{i,j=1}^{2} x_i x_j S_{ij} \quad (5)$$

$$\vec{Y} \cdot \vec{Y} = \sum_{i,j=1}^{2} y_i y_j \cos(\vec{a}_1,\vec{a}_2) = \sum_{i,j=1}^{2} y_i y_j S_{ij} \quad (6)$$

In N-dimensional Affine Space

If there are n Category Entities instead of two, we would have to use an n-dimensional affine space. For two points X and Y in n dimensional affine space A, the numbers x_i (i = 1, ..., n) and y_i (i = 1, ..., n) are the *affine coordinates* of X and Y with respect to the given coordinate system. The vectors $\vec{X}$ and $\vec{Y}$ have n sub-vectors in $\vec{a}_1, \vec{a}_2, \dots, \vec{a}_n$ directions. So for two vectors $\vec{X} = (x_1\vec{a}_1, x_2\vec{a}_2, \dots, x_n\vec{a}_n)$, $\vec{Y} = (y_1\vec{a}_1, y_2\vec{a}_2, \dots, y_n\vec{a}_n)$.

A vector addition rule called the Polygonal Law can add these sub-vectors with points from one end of a sub-vector to the other end.

$$\vec{X} = (x_1\vec{a}_1, x_2\vec{a}_2, \dots, x_n\vec{a}_n) = \sum_{i=1}^{n} x_i\vec{a}_i$$

$$\vec{Y} = (y_1\vec{a}_1, y_2\vec{a}_2, \dots, y_n\vec{a}_n) = \sum_{j=1}^{n} y_j\vec{a}_j$$

The Inner Product of vectors X and Y should be:

$$\vec{X} \cdot \vec{Y} = \sum_{i,j=1}^{n} x_i y_j \cos(\vec{a}_i,\vec{a}_j) = \sum_{i,j=1}^{n} x_i y_j S_{ij} \quad (7)$$

$$\vec{X} \cdot \vec{X} = |\vec{X}|^2 = \sum_{i,j=1}^{n} x_i x_j \cos(\vec{a}_i,\vec{a}_j) = \sum_{i,j=1}^{n} x_i x_j S_{ij} \quad (8)$$

$$\vec{Y} \cdot \vec{Y} = |\vec{Y}|^2 = \sum_{i,j=1}^{n} y_i y_j \cos(\vec{a}_i,\vec{a}_j) = \sum_{i,j=1}^{n} y_i y_j S_{ij} \quad (9)$$

In n Euclidean space (Cartesian coordinates), the included angle of vectors i and j is a right angle, so

$$S_{ij} = \cos(\vec{a}_i,\vec{a}_j), \begin{cases} i = j, S_{ii} = \cos(\vec{a}_i,\vec{a}_i) = 1 \\ i \neq j, S_{ij} = \cos(\vec{a}_i,\vec{a}_j) = 0 \end{cases}$$

$$\vec{X} \cdot \vec{Y} = \sum_{i,j=1}^{n} x_i y_j \cos(\vec{a}_i,\vec{a}_j) = \sum_{i=1}^{n} x_i y_i \quad (10)$$

$$\vec{X} \cdot \vec{X} = |\vec{X}|^2 = \sum_{i,j=1}^{n} x_i x_j \cos(\vec{a}_i,\vec{a}_j) = \sum_{i=1}^{n} x_i^2 \quad (11)$$

$$\vec{Y} \cdot \vec{Y} = |\vec{Y}|^2 = \sum_{i,j=1}^{n} y_i y_j \cos(\vec{a}_i,\vec{a}_j) = \sum_{i=1}^{n} y_i^2 \quad (12)$$

Concentration vs. Diversity in Affine Space

For one system, with the data of relative abundances of the Category Entities and the Similiarity among Category Entities, we can calculate Concentration or Diversity of a system.

Simpson Concentration Index and Gini-Simpson Diversity Index

We noticed that the formula (11) is the Simpson Concentration index. From the perspective of Geometry, Simpson Concentration Index depends on a distribution or occurrence frequency of these Category Entities in a system.

While the Gini-Simpson Index is a measure of diversity. From the formula (11), we get Gini–Simpson Diversity Index.

$$D(X,X) = 1 - \vec{X}\cdot\vec{X} = 1 - \left|\vec{X}\right|^2 = 1 - \sum_{i,j=1}^{n} x_i^2 \qquad (13)$$

Vector Norm in Affine Space and Rao-Stirling Diversity Index

We noticed that the formula (8) is squared vector norm in affine space, should be a Concentration Index. From the formula (8), we get Rao-Stirling Diversity. Ronald Rousseau gave the detailed algebraic rearranging to get formula (14) in Zhou et al (2012), also see Trezzini, B (2013).

$$D(X,X) = 1 - \vec{X}\cdot\vec{X} = 1 - \left|\vec{X}\right|^2 = 1 - \sum_{i,j=1}^{n} x_i x_j S_{ij} = \sum_{i,j=1}^{n} x_i x_j D_{ij} \qquad (14)$$

Where S_{ij}=1-D_{ij} is the similarity between Category Entities i and j.

From the perspective of Geometric, Gini-Simpson diversity index follow the rules in Euclidean space, while Rao-Stirling diversity index follow the rules in Affine space.

Similarity vs. Distance Measures in Affine Space

Inner Product in affine space

We find that co-occurrence matrices such are nothing more than inner products of vectors. In Scientometrics, co-occurrence matrices are often write in algebraic way, C_{xy} denote the co-occurrence matrix of the systems X and Y. C_{xy} is a symmetric non-negative matrix of order $n \times n$. For the non- diagonal of the symmetric matrix, the inner product usually write in the form $C_{xy} = \vec{X}\cdot\vec{Y} = \vec{X}^T\vec{Y}$, where $\vec{X}^T$ denotes the transpose of $\vec{X}$. For the diagonal of the symmetric matrix, the squared norms of the vectors can denoted by $\mathbf{C_x^2} = \vec{X}\cdot\vec{X}$ and $\mathbf{C_y^2} = \vec{Y}\cdot\vec{Y}$.

Leydesdorff and Vaughan (2006) insist that Similarity measures (such as cosine) should not be applied to the symmetrical co-citation matrix but can be applied to the asymmetrical citation matrix to derive the proximity matrix. In the opinion of Leydesdorff and Vaughan (2006), co-occurrence matrices are proximity data, which do not require conversion before mapping.

We prove from the perspective of Geometry, that co-occurrence matrices are already the similarity data (the inner product), so cannot use the cosine, but use Ochiai index to normalize. Many people mix up Cosine index and Ochiai index, actually, they are not the same thing.

The inner product is often at the heart of other similarity measures. In Cosine, Dice and Jaccard measures, the inner products act as the numerators. This division may be best understood as normalization. The critical difference, then, is the denominator of this formula (Jones et al, 1987).

The formula (7) for the inner product measure of similarity between two vectors in affine space is:

$$C_{xy} = \vec{X}\cdot\vec{Y} = \sum_{i,j=1}^{n} x_i y_j \cos(\vec{a}_i,\vec{a}_j) = \sum_{i,j=1}^{n} x_i y_j S_{ij}$$

With the inner product in affine space, we can generalize Cosine, Dice and Jaccard measures from Euclidean space to affine space.

The Vector Space Model (Cosine) in Affine Space

Cosine index

$$\cos(\vec{X},\vec{Y}) = \frac{\vec{X}\cdot\vec{Y}}{|\vec{X}||\vec{Y}|} = \frac{\vec{X}\cdot\vec{Y}}{\sqrt{\vec{X}\cdot\vec{X}}\sqrt{\vec{Y}\cdot\vec{Y}}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (15)$$

In affine space,

$$\vec{X}\cdot\vec{Y} = |X||Y|\cos_\theta$$

$$\cos\langle X, Y\rangle = \frac{\vec{X}\cdot\vec{Y}}{|X||Y|}$$

From formula (7), (8) and (9), we get, the similarity-weighted Cosine measure,

$$\cos(\vec{X},\vec{Y}) = \frac{\vec{X}\cdot\vec{Y}}{|\vec{X}||\vec{Y}|} = \frac{\vec{X}\cdot\vec{Y}}{\sqrt{\vec{X}\cdot\vec{X}}\sqrt{\vec{Y}\cdot\vec{Y}}} = \frac{\sum_{i,j=1}^{n} x_i y_j \cos(\vec{a}_i,\vec{a}_j)}{\sqrt{\sum_{i,j=1}^{n} x_i x_j \cos(\vec{a}_i,\vec{a}_j)}\sqrt{\sum_{i,j=1}^{n} y_i y_j \cos(\vec{a}_i,\vec{a}_j)}} \frac{\sum_{i,j=1}^{n} x_i y_j S_{ij}}{\sqrt{\sum_{i,j=1}^{n} x_i x_j S_{ij}}\sqrt{\sum_{i,j=1}^{n} y_i y_j S_{ij}}} \quad (16)$$

When getting the co-occurrence matrices (the inner product) directly from database, we can use the Ochiai index to normalize them.

$$O = \frac{C_{xy}}{\sqrt{C_x^2}\sqrt{C_y^2}} \quad (17)$$

Dice's Measure in Affine Space

For Dice's measure E in Euclidean space,

$$E = \frac{2\vec{X}\cdot\vec{Y}}{|\vec{X}|^2+|\vec{Y}|^2} = \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2+\sum_{i=1}^{n} y_i^2} \quad (18)$$

In affine space, Dice's measure should be,

$$Ea = \frac{2\vec{X}\cdot\vec{Y}}{|\vec{X}|^2+|\vec{Y}|^2} = \frac{2\sum_{i,j=1}^{n} x_i y_j \cos(\vec{a}_i,\vec{a}_j)}{\sum_{i,j=1}^{n} x_i x_j \cos(\vec{a}_i,\vec{a}_j)+\sum_{i,j=1}^{n} x_i x_j \cos(\vec{a}_i,\vec{a}_j)} \quad (19)$$

To normalize the co-occurrence matrix (the inner product), we can use the following form.

$$Ec = \frac{2C_{xy}}{C_x^2+C_y^2} \quad (20)$$

Jaccard and Tanimoto Index in Affine Space

$$J = \frac{\vec{X}\cdot\vec{Y}}{|\vec{X}|^2+|\vec{Y}|^2-\vec{X}\cdot\vec{Y}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2+\sum_{i=1}^{n} y_i^2-\sum_{i=1}^{n} x_i y_i} \quad (21)$$

$$Ja = \frac{\vec{X}\cdot\vec{Y}}{|\vec{X}|^2+|\vec{Y}|^2-\vec{X}\cdot\vec{Y}} = \frac{\sum_{i,j=1}^{n} x_i y_j \cos(\vec{a}_i,\vec{a}_j)}{\sum_{i,j=1}^{n} x_i x_j \cos(\vec{a}_i,\vec{a}_j)+\sum_{i,j=1}^{n} y_i y_j \cos(\vec{a}_i,\vec{a}_j)-\sum_{i,j=1}^{n} x_i y_j \cos(\vec{a}_i,\vec{a}_j)} \quad (22)$$

To normalize the co-occurrence matrix (the inner product), we can use the following form.

$$Jc = \frac{C_{xy}}{C_x^2+C_y^2-C_{xy}} \quad (23)$$

3 Application scenes

Systems can include, according to Ding (2013), evaluative entities (including documents, authors, journals, institutions, and/or countries) in bibliometric studies. Actually, they can be any system whose elements can be divided into categories.

The documents as Category Entities

The most common co-occurrence matrices, Bibliographic coupling, co-citation, co-word, and co-author matrices, are all the similarity based on documents (as Category Entities). The similarity matrix between documents can not obtained, so we assume the documents are independent. The similarity measure in Euclidean space, such as cosine, Jaccard index, Dice's index can be used in occurrence matrices to obtain those normalized co-occurrence matrices.

If we already get the row co-occurrence matrices, we can use the Ochiai index and other normalization index showed in formula (17), (20), (23).

Table 1 The Co-occurrence similarity relationship (Document as Category Entities)

	Systems				
	Word	Documents	Authors	Institute	Countries
Documents（articles，Patents）	Co-word	Co-citation, Bibliographic coupling	Co-author	Cooperation between Institute	Cooperation between Countries

The Knowledge Category Entities at different level

In bibliometric studies Ding (2013) defined Knowledge Entities as carriers of knowledge units in scientific articles which include such entities as keywords, topics, subject categories, datasets, key methods, key theories, and domain entities.

The similarity relationship between Knowledge Category Entities can be obtained (see Table 2), such as we can obtain semantic relation between terms from Mesh, HowNet, WordNet, et al. Similarity among Subject Categories based on the citation relationship. IPC hierarchical taxonomies tree also indicate the "is a" relationship between IPC category Entities.

Table 2 The similarity relationship of different Knowledge Category Entities

Knowledge Category Entities	Relationship	Data resouces	Similarity matrix
Terms	Semantic relation	Ontology、Thesaurus（e.g. Mesh,HowNet,WordNet）	Similarity matrix
Patent category	Technology distance relation	IPC hierarchical taxonomies	IPC Similarity matrix
Subject category	Interdisciplinary	Subject Categories Citation Matrix	Subject Categories Citation Matrix

With the similarity relationship of different Knowledge Category Entities, and the occurrence matrix of Knowledge Category Entities as Columns and different systems as Rows in Scientometrics, we can caudate the diversity and similarity measure in affine space with the formula (8), (16), (19), (22).

Table 3 some occurrence matrices in Scientometrics (Knowledge Category Entities at different level)

Knowledge Category Entities (Column)	Systems (Rows)			
	Document	Author	Institute	Country
Terms	Document similarity based on terms	**Author similarity based on Keyword**	Institute similarity based on terms	Country similarity based on terms
Patent Category	Document similarity based on Patent Category	Author similarity based on Patent Category	**Company similarity based on Patent Category**	Country similarity based on Patent Category
Subject Category	Document similarity based on Subject	Author similarity based on Subject Category	Institute similarity based on Subject Category	**Country similarity based on Subject Category**

	Category			

4 Empirical studies

Three Empirical studies are made to show the result of Knowledge Category Entities at different level.

The homogeneity of disciplinary structure between countries

Similarity or dissimilarity of Subject Categories is taken into account, when computing the Subject Category diversity of a country and similarity among countries. The homogeneity of disciplinary structure between countries based on Subject Categories is measured using the similarity-weighted cosine measure. This work has been published in Zhou et al (2012).

Technological distance between firms

The dissimilarity matrix represented the taxonomic distinctness in an IPC hierarchical tree. The occurrence matrix consists of the patents IPC proportion of 12 companies downloaded from DII database. With the two matrices, the dissimilarity-weighted distance is utilized to measure the technological distance between different firms. A paper is preparing according to this work.

The Author similarity based on MeSH

The Author-MeSH occurrence matrix and the similarity matrix (is-a relationship) of MeSH are used to measure the author similarity. This work is in progress.

Motivation for student participation at the Doctoral Forum and the issues you wish to receive feedback on from the senior researchers.
I want to know is my theoretical derivation of analytic geometry reasonable? And I welcome any comments.