Similarity Measures in Documents Using Association Graphs

José E. Medina Pagola[1], Ernesto Guevara Martínez[2], José Hernández Palancar[1], Abdel Hechavarría Díaz[1], and Raudel Hernández León[1]

[1] Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), 7ª # 21812 e/ 218 y 228, Rpto. Siboney, CP. 12200, Playa, C. de la Habana, Cuba
{jmedina, jpalancar}@cenatav.co.cu
[2] Instituto Superior Politécnico "José Antonio Echeverria" (ISPJAE), Ave. 114 # 11901, CP. 10390, Marianao, C. de la Habana, Cuba
eguevara@ceis.cujae.edu.cu

Abstract. In this paper we present a new model, designated as Association Graph, to improve document representation, facilitating the ontological dimension. We explain how to generate and use this kind of graph. Also, we analyze different document similarity measures based on this representation. A classical vector space model was used to evaluate this model and measures, investigating their strengths and weaknesses. The proposed model was found to give promising results.

1 Introduction

At the moment, due to vertiginous scientific and technological advances of the last years, institutions have great capacities of creating, storing and distributing their data. This situation, among other things, has increased the necessity of new tools that aid in transforming this vast quantity of data in useful information or new knowledge that can be used in decision making. Data mining systems are examples of this type of tools.

These systems allow us to analyze and to discover interesting patterns in large databases. However, due to the information characteristics contained in traditional databases and data warehouses, data mining systems are not appropriate for the analysis of other types of information less structured like, for example, the one contained in text collections. For this reason, Text Mining arises as an alternative to understand the processing of natural language. Text Mining combines artificial intelligence, statistical, database, and graphic visualization techniques, allowing the comprehension of aspects dealing with the identification, organization and understanding of the knowledge appearing in any text.

Examples of systems that use those techniques, and have gotten some attention in recent years, are pointed out by Yao et al. as RSS (Research Support Systems) and WRSS (Web-based RSS) [1]. They improve current search tools, helping scientists to access, explore, evaluate and use information on digital libraries or on the Web, improving research productivity and quality [2].

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 741 – 751, 2005.

Text Mining, together with others techniques, such as profiling, collaborative filtering, intelligent agent, etc., should be considered to develop those systems. Text Mining, as many other tasks of text processing, is usually carried out on simple representations of text contents. However, profiling, collaborative filtering and WRSS require more complex semantic relations, usually expressed as semantic graphs [3].

In this paper we propose an approach using Association Graphs, a measure as an alternative representation of documents and a way of measuring their similarities, facilitating their ontological dimensions required by many applications as, for instance, WRSS. In Section 2 we will present general considerations for vector space models in Text Mining. In Section 3 we will analyze the limitations of term correlation for knowledge indexing and representation. In Section 4 we will explain our proposal, as an alternative to improve document representation, facilitating the ontological dimension.

2 Text Mining

Text Mining could be defined as a discovery process of interesting patterns and new knowledge in a text collection; therefore, Text Mining is a specific type of Data Mining applied to documents to discover information not present in any specific one. Hence, its objective is to discover things such as regularities, tendencies, deviations and associations in huge databases in textual form [4].

By applying algorithms of Text Mining to documents stored in different media, for example in WRSS, one may discover patterns and extract knowledge useful to decision-makers, in the example researchers, who are interested in exploratory searching and browsing [1].

The process of Text Mining is carried out in two main stages: a pre-processing stage and a discovery stage. In the first stage, texts are transformed into a kind of structured or semi-structured representation, facilitating their later analysis. In the second stage these representations are analyzed in order to discover interesting patterns or new knowledge [4].

In the pre-processing stage a set of operations is done to simplify and standardize the texts being analyzed. Some of the operations considered are the following:

- Recognizing useful words.
- Ignoring the null words, also known as Stopwords.
- Identifying phrases or terms with multi-words.
- Obtaining the canonical forms of the words, also known as stemming.

As a result of this stage a sequence of distinguished terms is obtained. These terms could be organized in different forms but, in general, they are considered as groups or bags of terms, usually structured using vector models [5]. In these representations, the sequences of the terms, their correlations or syntactical relations are not analyzed; therefore, their mutual independence is supposed. The values of those vectors could be assumed as weights, considering the following interpretations [6]:

- Boolean - Each term is associated with a Boolean value representing if it is present or not in a document.
- TF (Term Frequecy) - Each term is associated with a frequency of appearance in a document, absolute or normalized.
- TF-IDF (Term Frequency - Inverse Document Frequency) - The term is associated with its frequency, adjusted by the inverse of the number of documents containing each term.

These vectors of terms are used in a second stage, among other tasks, to analyze the similarities between documents, or groups of them, using different measures as the Cosine, applied to the angle between the vectors, define as [6]:

$$sim(d_i, d_j) = \cos(d_i, d_j) = \frac{(d_i \bullet d_j)}{\| d_i \| * \| d_j \|} = \frac{\sum w_{ir} * w_{jr}}{\sqrt{\sum {w_{ir}}^2 * \sum {w_{jr}}^2}} \ , \quad (1)$$

where d_i, d_j are the vectors of documents i, j, $\|d_i\|$, $\|d_j\|$ the norms of the vectors, and w_{ir}, w_{jr} are the term weights in the vectors d_i, d_j, respectively.

3 Ontological Requirements

Although, generally, the terms appearing in a document are interrelated and the vector space model, proposed by Salton [7], has been the dominant way to represent and measure document similarities, some authors consider this treatment as an elementary way of the ontological dimension of the information.

While that treatment could be adequate for some applications, in others, like WRSS and collaborative filtering systems, more complex semantic relations are required. Collaborative filtering system, a kind of information filtering, evaluates resources in order to recommend objects preferred by similar users, supposing they are also useful to a particular user [8].

In WRSS, documents are the resources to be evaluated. In this case, the scientific knowledge of documents, or groups of them, and scientific profiles of users should be considered. That knowledge and profiles are usually expressed by semantic graphs constructed generally by users. For that reason, one should evaluate methods for automatic or semi-automatic graph generation, quite difficult to make from a simple vector model.

An alternative approach of the ontological dimension is observed in [9]. In this work the authors use Conceptual Maps to identify potential terms and relationships. So, with this proposal, the user defines his personal Conceptual Maps interactively. Although the author's intention might be the use of Conceptual Maps in an information retrieval process, such approach wasn't discussed in that work.

Other ways to include an ontological dimension are the corpus-based methods in conjunction with lexical taxonomies to calculate semantic similarity between words/concepts. Examples of these methods are those developed over the broad-coverage taxonomy known as Wordnet [10].

Well alternative approaches to the vector space model are the language models. These consider the probabilities of occurrence of a phrase *S* in a language *M*, indicated by *P(S/M)*. However, the phrases are usually reduced to one term, assuming again unigrams and independence among them. An example of this model is the Kullback-Leibler Divergence (a variation of the cross-entropy), defined as:

$$D(d_i \parallel d_j) = \sum P(t / d_i) \log \frac{P(t / d_i)}{P(t / d_j)}.$$

This expression could be combined in both directions to obtain a similarity measure, as was pointed out by Feldman and Dagan [11].

An interesting implementation is the proposal of Kou and Gardarin [12]. This proposal is a kind of language model, considering the similarities between two documents as:

$$sim(d_i, d_j) = d_i \bullet d_j = \sum_r w_{ir} w_{jr} + \sum_r \sum_{s \neq r} w_{ir} w_{js} (t_r \bullet t_s),$$

where w_{ir} and w_{js}, using Kou-Gardarin terminology, are the term weights in document vectors d_i, d_j, respectively, and $(t_r \bullet t_s)$ is the a priori correlation between terms t_r and t_s. Actually, the authors included in the first part of the expression the self-correlation in t_r, considering that $(t_r \bullet t_r) = 1$. The authors propose the estimation of the correlation through a training process. As can be noticed, those correlations express the probabilities $P(t_r,t_s/M)$ of phrases containing the terms t_r, t_s in a language *M*. Besides, that expression could be reduced to the Cosine measure (normalized by the length of the vectors) if the term independence is considered and, for that reason, the correlation $(t_r \bullet t_s)$ is zero.

Although the Kou-Gardarin proposal improves the independence limitation of the vector space model, it considers that two terms are correlated as a tendency, and independent of the documents analyzed in the similarity measure. This assumption underestimates the ontological view of each document.

The approaches mention above are variants of the Generalized Vector Space Model proposed by S.K.M Wong et al. [13]. In their work, they expressed that there was no satisfactory way of computing term correlations based on automatic indexing scheme.

We believe that up to the present time that limitation has not been solved yet. Although several authors have proposed different methods of recognizing term correlations in the retrieval process, those methods try to model the ontological dimension by a global distribution of terms, but not with a local evaluation of documents.

In general, it could be assumed that with a better ontological representation of the information retrieved and discriminated, the better the documents will be mined. Besides, it is expected that a better representation improve the capacity of knowledge comprehension regarding the vector model. These considerations will be developed in more details later on.

4 Association Graphs

It is comprehensible that a same term in two documents could designate different concepts. Besides, two terms could have different relations, according to the subject of each document, and those relations could exist only in the context of some documents, forming a specific group, and independent of the relations in a global dimension or language.

In order to model the relation between two terms in a document, we will consider the shortest physical distance between those terms. So, two documents shall be closer if the number of common terms is greater and the shortest physical distances among those terms are similar. With these assumption we hypothesize that, in order to recognize the semantic relation between two terms, it is enough that they appear together at least once in a small context: a sentence, a paragraph, and so on.

The use of physical distance among terms has been considered in other works. For example, Ahonen et al. has appointed that many documents, especially books and papers, are structured in sections or micro-documents and, logically, terms in a same micro-document are strongly related, but in different micro-documents the physical relation uses to be weak [14]. Although they realized the relevance of the physical relation among terms, the vector model was considered in their work.

Also, many search engines to measure the document's importance or quality consider the proximity among the words of complex equations or queries.

In order to measure the distance between two terms t_r and t_s in a document i, designated by D^i_{rs}, the physical distance in the document between those terms could be defined in different ways. One way could be considering the number of words between them. Although this could be a feasible solution, it ignores the semantic strength in sentences and paragraphs.

Considering the distance by sentence, D^i_{rs} will be $n+1$, where n is the number of intermediate sentences between those containing the terms..

If we consider the distance by paragraph, without ignoring the natural co-occurrence when appearing in the same sentence, and considering: (p_r, n_r), (p_s, n_s), the paragraph and sentence numbers of terms t_r and t_s respectively, the physical distance between these terms is defined as follows:

$$D^i_{rs} = \begin{cases} 1 & (r = s) \vee [(p_r = p_s) \wedge (n_r = n_s)] \\ |p_r - p_s| + 2 & \text{Other case} \end{cases}.$$

Observe that the minimum value of D^i_{rs}, as could be expected, isn't zero, but one in both cases.. This consideration is only a convenient assumption to expressions defined farther on.

Besides, it will be considered in both distance that every term is related to itself, having distance one, in order to include the case two documents have only one term in common.

According to this, a document could be modeled by a graph, where the nodes are the distinguished terms and the arcs are their relations, weighted by their distances. Also, we are considering this is a full connected graph, having any term some relation (stronger or not according to the distance) with the others.

Although the physical relation, in conjunction with the common terms, could be used to evaluate the neighborhood among documents, the weights of the distinguished terms should not be ignored in a similarity measure. To include these values, the document graph could be extended with weighted nodes.

Therefore, a first approximation for a document representation could be seen as a weighted graph by node, considering the weights of the distinguished terms, and by arc, considering the shortest physical distance between the adjacent terms.

As the additional components of these graphs are the arcs, with respect to the vector model, and trying to combine the weights of the terms and the distance between them to express the strength of their association, the vector A^i_{rs}, named *Association Vector*, is proposed as the arc's weight of the related terms t_r, t_s in a document i, defined as:

$$A^i_{rs} = \left(\frac{w^i_r}{\sqrt{D^i_{rs}}}, \frac{w^i_s}{\sqrt{D^i_{rs}}}\right) ,$$

where w^i_r and w^i_s are the weights of the terms t_r and t_s, respectively, in a document i.

As the arc's weight A^i_{rs} is a two-dimensional vector, the strength of the terms's association can be evaluated as the Euclidean norm $\|A^i_{rs}\|$. In this case, the strength is greater if the terms's weights are greater and the distance between the terms is shorter. Besides, the upper value of A^i_{rs} is (w^i_r, w^i_s), when the distance is one, and the lower tends to (0, 0), when the distance is very long.

With these transformations, an *Association Graph* can be defined as a weighted graph by arc, considering as weight of each arc (t_r, t_s) the Association Vector A^i_{rs}.

5 Similarity Measures

Although for a vector model a Cosine measure represents a standard way to evaluate the similarity between two documents, in a graph model (as the Association Graph) other measures should be considered.

As our graph doesn't posses a structural or spatial representation, it is enough to treat it as a set of arcs. Several authors have proposed different matching coefficients for sets, which in general coincide with commonly used measures of association in information retrieval.. Examples of these are: Dice's, Jaccard's and Overlap coefficients, among others [15]. These may all be considered to be normalized versions of the simple matching coefficient of two sets X and Y, defined as: $|X \cap Y|$.

Another version of the simple matching coefficient is the proposal of Pazienza and Vindigni. They define a common coverage of two non-empty sets as the average of the coverage of their intersection with respect to each of them [16].

As we are considering sets of arcs, a first idea for a matching coefficient is trying to define a simple matching-like one. If that were adequate for a common graph, in an Association Graph, where each graph has different association strengths, the coefficient could be better constructed as the Pazienza-Vindigni proposal.

According to the previous idea, and considering the Association Graphs of documents i, j, the *Simple Coverage* as a similarity measure could be used, expressed as:

$$sim(d_i, d_j) = \tfrac{1}{2} \frac{\sum_{t_r, t_s \in T_{ij}} \| A^i_{rs} \|}{\sum_{t_r, t_s \in T_i} \| A^i_{rs} \|} + \tfrac{1}{2} \frac{\sum_{t_r, t_s \in T_{ij}} \| A^j_{rs} \|}{\sum_{t_r, t_s \in T_j} \| A^j_{rs} \|} \quad , \tag{2}$$

where T_i, T_j represent the sets of terms in the Association Graphs of documents i, j, respectively, and T_{ij} is the set of the common terms ($T_i \cap T_j$).

Notice that the first part of the expression evaluates the proportion of the total association strength of the common arcs with respect to the total strength in whole document i, and the second part the same but in document j. The fractions ½ in the formula guarantee that this measure has values in the interval [0, 1].

Although we considered that the Equation 2 is a good first approach, we realized that it doesn't measure the similarities between the vectors associated with the common arcs. In order to include these similarities, we propose the *Weighted Coverage* measure, defined as:

$$sim(d_i, d_j) = \tfrac{1}{2} \frac{\sum_{t_r, t_s \in T_{ij}} S^{ij}{}_{rs} \| A^i_{rs} \|}{\sum_{t_r, t_s \in T_i} \| A^i_{rs} \|} + \tfrac{1}{2} \frac{\sum_{t_r, t_s \in T_{ij}} S^{ij}{}_{rs} \| A^j_{rs} \|}{\sum_{t_r, t_s \in T_j} \| A^j_{rs} \|} \quad . \tag{3}$$

If T_i or T_j are empty sets, the expression is defined as zero. The weight $S^{ij}{}_{rs}$ represents a similarity measure between $A^i{}_{rs}$ and $A^j{}_{rs}$. This weight is defined in this paper as:

$$S^{ij}{}_{rs} = \cos(A^i_{rs}, A^j_{rs}) * (1 - \tfrac{1}{2}(\| A^i_{rs} \| - \| A^j_{rs} \|)^2) \, ,$$

where the first part of the expression represents the cosine between those vectors, defined in a similar way as the Equation 1. It can be noticed that the weights defined in this manner include not only the angles between the vectors, but also the differences of their strengths.

This similarity measure could be extended to evaluate the similarities between documents, groups of documents, and user profiles, changing the values ½ of each part of the formula by different fractions. These extensions could be convenient to many applications, as collaborative filtering and WRSS.

6 Experiment and Analysis

In order to evaluate the proposed measure, the data TREC-5 in Spanish (http://trec.nist.gov) was used. From this data, we used 676 news published by AFP during 1994 and classified in 22 topics. Table 1 shows the topics and the quantity of documents for each topic in this data.

The pre-processing stage was done with the library of the system JERARTOP [6], which used the morphological analyzer MACO+, developed by the Natural Language Processing Group of the Polytechnic University of Catalunya, based on extended

Table 1. Topics of TREC-5

Topic	Description	# Doc.
SP51	Ocean's Fish Suplí	75
SP52	Basque Rebels War	13
SP53	Women's status in Latin America	46
SP54	World's Marine Resources	35
SP55	Fate of Carlos Andrés Pérez	108
SP58	Financing of Samper Election	44
SP59	Hoof and Mouth Disease	7
SP60	Methods Narcotraffickers Use to Hide their Drugs	46
SP62	Colombia's Fresh Flower Trade	15
SP63	Drug Trafficking Involvement	5
SP64	Green Iguana Extinction	7
SP65	Raul Castro's Activities	29
SP66	MERCOSUR	68
SP67	Peruvian Fishmeal Industry	8
SP68	AIDS in Argentina	12
SP69	Status of Russian Satellites and Membership in NATO	62
SP70	NATO Peace Force in Bosnia	6
SP71	Status of United States' Certification of Columbia and its War on Drugs	11
SP72	Damage to Mexico's Environment	14
SP73	Illegal Trade of Exotic Animals	12
SP74	Privatization of Major Sectors of Argentina Economy	34
SP75	Heroin in Latin America	19
Total	22 Topics	676

stochastic models ECGI [17]. A detailed description of that analyzer can be found in http://www.lsi.upc.es/~nlp.

A classical vector model was used to evaluate the proposed approach, applying the Cosine measure. The term weights were calculated as TF (*Term Frequency*), normalized by the maximum frequency. K-Nearest Neighbour classifier, with weighted voting by similarity value, was conducted by taking the value of K as 5, 10, 15 and 20. A k-fold cross-validation was applied with k=10. The results obtained are shown in Table 2, where *simC* and *simG* are the measures obtained by Cosine and Weighted Coverage models respectively.

Precision, Recall and F1 are three commonly used evaluation measures of performance. For a single category or topic, these measures can be defined as [18]:

Precision = "Correctly assigned" / "Assigned to the category"
Recall = "Correctly assigned" / "Belonging to the category"
F1 = 2 * Recall*Precision / (Recall+Precision)

Table 2. Macro-averaged Performance

K	Precision		Recall		F1	
	simC	*simG*	*simC*	*simG*	*simC*	*simG*
5	0.8175	0.8222	0.7289	0.7356	0.7545	0.7611
10	0.8072	0.8568	0.6663	0.7088	0.6973	0.7414
15	0.8566	0.8482	0.6452	0.7043	0.7079	0.7397
20	0.8425	0.8461	0.6233	0.6813	0.6836	0.7181

Precision, Recall and F1 are three commonly used evaluation measures of performance. For a single category or topic, these measures can be defined as [18]:

Precision = "Correctly assigned" / "Assigned to the category"
Recall = "Correctly assigned" / "Belonging to the category"
F1 = 2 * Recall*Precision / (Recall+Precision)

For evaluating the performance average across categories, there are two conventional methods: Macro-averaging performance and Micro-averaging performance. Macro-averaged performance scores are computed by a simple average of the performance measures for each category. Micro-averaged performance scores are computed by first accumulating the corresponding variables in the per-category expressions, and then using those global quantities to compute the scores. Micro-averaged performance score gives equal weights to every document. Likewise, macro-averaged performance score gives equal weights to every category or topic, regardless of its frequency.

As can be noticed in Table 2, Association Graph model outperforms Cosine similarity model for different K values, except for Macro-precision with K=15. Besides, as an average, 2.9 % of F1 measure in Weighted Coverage model is bigger than in Cosine model. This proves that the use of physical term association really improves the effectiveness of categorization.

Although these results are only preliminaries, they show that the Association Graph and the proposed measure represent a good model and seem to be better than the Vector-Cosine.

7 Conclusions

Although some approaches have been considered, especially in semi-automatic processing, the vector space model has been the dominant way for document representations, especially as frequency vectors of terms. These representations are relatively easy to build from texts, but cannot express several details of their meanings, having a poor capacity of description. In order to achieve a better representation of the knowledge contained in documents, we have proposed the Association Graphs.

Using this kind of graph, a similarity measure, named Weighted Coverage, is proposed, making it possible to compare and discriminate documents, applying it in different techniques as, for example, clustering and classification algorithms.

Some variations to the proposed measure could be analyzed and other distance measures could be assumed as, for example, limiting the distance to a convenient value.

Nevertheless, the experiment has shown interesting results. Although other experiments must be done, the proposed model was found to give promising results.

References

1. Yao J.T., Yao Y.Y.: Web-based Information Retrieval Support Systems: building research tools for scientists in the new information age. Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence, (WI 2003), Halifax, Canada (2003).
2. Xu J., Huang Y., Madey G.: A Research Support System Framework for Web Data Mining. Proceedings of WI/IAT 2003 Workshop on Applications, Products and Services of Web-based Support Systems, WSS 2003, Halifax, Canada (2003).
3. Rojo A.: RA, un agente recomendador de recursos digitales de la Web. Master thesis, Universidad de las Américas, Puebla, México (2002). URL: http://www.pue .udlap .mx / ~ tesis/msp/rojo_g_a/.
4. Berry M.: Survey of Text Mining, Clustering, Classification and Retrieval. Springer (2004).
5. Raghavan V., Wong S.: A critical analysis of Vector Space Model for Information Retrieval. Journal of the American Society on Information Science, Vol. 37, No. 5, pp. 279-287 (1986).
6. Pons A.: Desarrollo de algoritmos para la estructuración dinámica de información y su aplicación a la detección de sucesos. Doctoral thesis, University Jaume I, Spain (2004).
7. Salton, G., The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs, New Jersey (1971).
8. Ziqiang W., Boqin F.: Collaborative Filtering Algorithm Based on Mutual Information. APWeb 2004, LNCS 3007, pp. 405-415. Springer-Verlag Berlin Heidelberg (2004).
9. Simón A., Rosete A., Panucia K., Ortiz A.: Aproximación a un método para la representación en Mapas Conceptuales del conocimiento almacenado en textos, con beneficios para la Minería de Texto. I Simposio Cubano de Inteligencia Artificial, Convención Informática 2004, Cuba (2004).
10. Budanitsky A., Hirst G.: Semantic distance inWordNet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000) (2001).
11. Feldman R., Dagan I.: Knowledge Discovery in Textual Databases (KDT). Proceedings of the first International Conference on Data Mining and Knowledge Discovery, KDD'95, Montreal, pp. 112-117 (1995).
12. Kou H., Gardarin G.: Similarity Model and Term Association for Document Categorization. NLDB 2002, LNCS 2553, pp. 223-229, Springer-Verlag Berlin Heidelberg (2002).
13. Wong S.K.M, Ziarko W. and Wong P.C.N.: Generalized Vector Space Model in Information Retrieval. Proc. of the 8th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, New York, ACM 11 (1985).
14. Ahonen H., Heikkinen B., Heinonen O., Klemettinen M.: Discovery of Reasonably sized Fragments Using Inter-paragraph Similarities. Technical Report C-1997-67, University of Helsinki, Department of Computer Science (1997).
15. C. J. van Rijsbergen C.J.: Information Retrieval. London: Butterworths (1979).

16. Pazienza M.T. and Vindigni M.: Agents Based Ontological Mediation in IE Systems. SCIE 2002, LNAI 2700, Springer-Verlag Berlin Heidelberg (2003).
17. Carmona J., et al.: An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98 (1998).
18. Yang Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, Vol. 1, No. 1/2, pp. 67-88 (1999).