ADCS'OO Keynote Address

Web Wombat - Building the Premier Australian Search Engine

Phillip Betolus
Technical Director
Web Wombat
www.webwombat.com

December 1,2000

Abstract

The Web Wombat story from a technical perspective (excluding the nitty gritty secret stuff). Starting at the origin of the Web Wombat search engine concept and the initial engines, developed in the back room of the house, through to search engines running from various parts of the world. There are also some forward looking statements about the future of search, which may or may not come to pass and do not reflect any particular business plans Web Wombat may or may not have.

About the Speaker

Phill Bertolus is Technical Director and cofounder of Web Wombat. Phill has 20 years experience in computing environments as diverse as aviation, banking, military. and Internet industries and has written about telecommunications and mobile phones for Australia's leading newspapers (The Age. The Sydney Morning Herald and the Australian Financial Review).
He wrote a compiler by the age of 23 and has successfully developed and internationally marketed a number of PC products.
His special areas of interest include compiler design, search engine design and realtime computing. Phill is a computer engineering specialist with a BE (Electronic Engineering) and a DipE (Communications Engineering).

Towards an Efficient Retrieval of Medical Imaging

Richard CHBEIR *Franck FAVETTA*

Laboratoire d'Ingénierie des Systèmes d'Information (LISI)
20 av Einstein, F-69621 Villeurbanne - France
Phone: (+33) 4 72 43 88 99 - Fax: (+33) 4 72 43 87 19

richard.chbeir@ieee.org franck.favetta@ ieee.org

Abstract

Image description is not an easy task. The same image can be described through different views: on the basis of either low-level properties, such as texture or color; context, such as date of acquisition or author: or semantic content, such as real-world objects and relations. Our approach consists in providing a global description solution capable of integrating different dimensions (or views) of a medical image. Via our approach, we are able to propose a solution that takes into consideration the heterogeneity of user competence (physician, researcher, student, etc.) and a high expressive power for medical imaging description. Visual solutions are recommended and are the most suited for non "novice" users in computing. However, current visual languages suffer from several problems as imprecision and no respect of integrity of spatial relations. Particularly, resolution of ambiguities generated by the user and/or the system at different levels of image description remains a challenge. In this paper, we present our solution for resolving these issues. A prototype has been implemented.

Keywords: Information Retrieval, Medical Imaging, Spatial Relations, Ambiguity Resolving.

1 Introduction

Image description is person-dependent and confined to parameters such as context, image objects, application domain, etc. The current approaches take into consideration only certain image facets. In a historical museum, image retrieval is based on the semantic content linked to objects and the scene (context) relating them. In monitoring applications, the searching process is built upon dates (all images at 13/12/1999). shapes, colour and texture.

Proceedings of the 5th Australasian Document Computing Symposium, Twin Waters Resort, Sunshine Coast, Australia, December 1, 2000.

We have found that current approaches suffer from different weaknesses:

1- Incomplete description of image content: an interpretation of a satellite image of the moon may ignore other stars.
2- Ambiguities resulting from the incompatibility between system pertinence and user relevance.
3- Impossibility of reutilization: a database of medical x-rays in hospital "A", for instance, cannot be reused in hospital "B" where physicians exercise another specialty because acquisition and retrieval systems are domain-dependent.

A recent study has begun in our laboratory (**LISI**) [9, 10, 11] aiming to provide an adequate image retrieval solution for heterogeneous users in the medical domain. We propose a solution built on **global description** process at storage process which takes into account complexity and variety of medical imaging (MRI, Scanner, X-rays, etc.). If image description during storage process takes into account more elements and different points of view, retrieval process becomes efficient for global and various types of queries. In this paper, we present in particular how ambiguities formulated by the user during query process can be resolved.

The next section details related work in this domain. Section 3 presents our data model. Section 4 presents the description process used by our system. Section 5 shows ambiguities problems and solutions. The last section concludes and summarizes our future directions.

2 Related work

In the literature, the description problem of images has been accessed through different approaches:

- **The contextual approach** considers only external data such as acquisition date, author name, file name or artificial keys [23]. However, this is restrictive and inappropriate for several domains such as Medical Imaging, TV production, multimedia, art history, geology, satellite image databases, etc.
- **The abstraction approach** manipulates the physical (low-level) properties of an image such as color histogram [4], texture [6, 12], shape [20, 7], edge [19, 21], etc. The abstraction approach is applied to domains where various types of images exist such as TV databases, museum databases, etc. This approach is currently having great interests because of automatic procedures used to compare images. However, abstraction approach is not appropriate in medical domain because of the diversity of the human body and the time-consuming procedures required for each digitalization techniques (Scanner, **MRI,** X-ray, etc.).
- **The classification approach** treats high-level properties of the image and describes its semantic content in terms of real objects and relationships. Object description and classification are generally done manually (or semi-automatically) at image storage. Describing the content and the sense of each object is difficult because probable descriptions are numerous and each person may describe the image differently. Secondly, image description based on object position and relationships between objects (spatial facet) has proven to be imperfect at retrieval process where translation, scaling, perfect and multiple rotations, or any arbitrary combination of transformations is applied. Thirdly, a great waste of data is induced when replacing an image by a set of poor semantic descriptors to describe image objects.

As the main objective of our project consists in providing retrieval medical imaging destined for non-professional end-users, user-friendly languages are required. Several language types have been proposed the last decades [2, 16, 24]. Current retrieval systems proposed in the medical domain [1, 13, 25] are not user-oriented and only few query possibilities are proposed. Visual languages can be seen as precursors in many domains applications. They aim at supplying user-friendly interfaces, especially for handling of spatial criteria in spatial queries. Several approaches have been proposed [5, 8, 17, 22], presenting advantages and drawbacks in terms of ambiguity and user-friendliness [18]. The main advantage of visual languages comes from the fact that the user does not have any constraint to express a query and no new language to learn. Nevertheless, many limitations still remain. The main limitation conies from the ambiguities of visual languages.

3 Data model

Image content is very rich in terms of properties, characteristics. salient objects and relationships. Fundamentally, the medical image content represents an aggregation of **salient internal** image objects: *organs, atomic regions,* and *anomalies.* Each salient object possesses low-level features (texture, shape, color, etc.) and high-level features (semantic noun, definition, synonyms, etc.). Based upon Eakin's framework [14], our model integrates four dimensions:

- The **contextual dimension** regroups global and external image properties (external object) without taking image's content into account.
- Each salient objects can be mapped into three orthogonal dimensions:
  - The **physical dimension** regroups local physical properties of image content as colors, textures and other low-level features.
  - The **spatial dimension** takes into consideration shapes (polygons) and spatial relations (cardinality and topology) calculated between objects.
  - The **semantic dimension** integrates semantic objects and relations.

4 Description process

To support our proposition, an implementation called MIMS (Medical Image Management System) has been realized [9, 10, 11]. The description of the whole system is out of scope of this paper. The architecture of MIMS is surveyed here below, followed by a short exposition of the topological precision of MIMS, and a description of the *ambiguity resolver* component.

4.1 Architecture

MIMS is composed of several components (Figure 1): analyzer, ambiguity resolver, SQL Translator, and Spatial Knowledge Model (SKM). Each component has its own task. In short, the analyzer assists the user to describe the image. The ambiguity resolver checks ambiguities found in each dimension (semantic, spatial, and physical). The SQL translator transforms commands into SQL statements. The Spatial Knowledge Model, based on Spatial Knowledge Base and the Inference Base, plays the consultant task and guides the ambiguity resolver to decide, whenever an ambiguity is found, whether the system or the user is able to respond. The analyzer, in collaboration with the Spatial Knowledge Base (a component of SKM), is responsible for providing possible features to describe image dimensions (spatial, physical, semantic, and contextual). In essence, the spatial knowledge base (SKB) is a set of structures (spatial, physical, hierarchical, semantic and evolutive) that embed different medical image dimensions.

MIMS analyzer passes by several steps:

- Contextual analysis: this step is used to acknowledge the contextual dimension.
- Physical analysis: this step attempts to describe physical dimension of medical image.
- Spatial analysis: after physical analysis, objects' (and regions') locations and positions can be calculated. MIMS calculates two kind of spatial relations: *directional relations* describing *cardinality* (North, East, Left, Bottom, etc.) and *topological relations*. The next subsection describes the topological expressive power of MIMS.

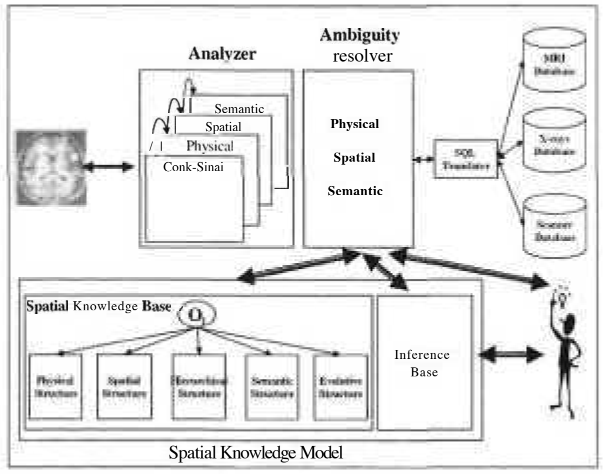

Figure 1: MIMS Architecture.

- Semantic analysis: treats semantic terms (objects and relations) proposed by the user during description process. Spatial Knowledge Model verification is applied during semantic dimension identification.

After semantic analysis, the system sends features to the SQL translator. The latter transforms features translated by ambiguity resolver into SQL statements and sends them to the JDBC[1] driver. In its turn, the JDBC driver examines SQL statements and accesses relative databases (MRI, X-rays, etc.). *Figure 2* shows the analyzer's results of a description process during the storage process.

[1] Java DataBase Connection (JDBC): permits to connect and remote heterogeneous databases using an API supplied in Java Development Kit.

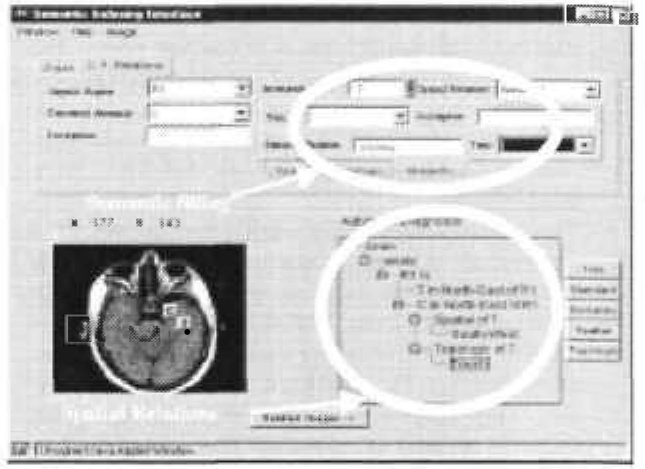

Figure 2: Screen shot of storage interface where image description is realized.

4.2 The topological precision

Topological relations constitute one of the fundamental concepts necessary for the description of spatial data. They are preserved under topological transformations such as translation, scaling, and zooming. The *9-Intersections* model [15] gives a formal categorization of topological relations involving two regions without holes in R_2 *(Figure 3). A* new model, based upon the previous model, gives a formal categorization of all the possible topological relations with three spatial objects in R_2: the *20-Intersections* model (Figure 5).

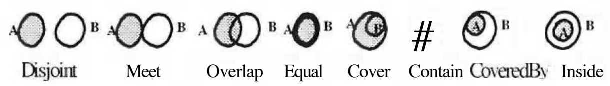

Figure 3: The *9-Intersections* model [15] categorizes eight topological relations between two regions in R_2.

A set of objects with defined *binary* relations can have several configurations which are not *differentiated*, e.g., the two configurations of Figure 5 both have the same *binary* topological description <a1 Meet a2> ^ <a2 Meet a3> ^ <a3 Meet a1> but they are different. To make this distinction is important, especially for the medical domain which requires a high topological expressive power, e.g., the difference between two configurations for the three anomalies of Figure 5 can be crucial for the physician. The *20-Intersections* model provides such a higher topological expressive power.

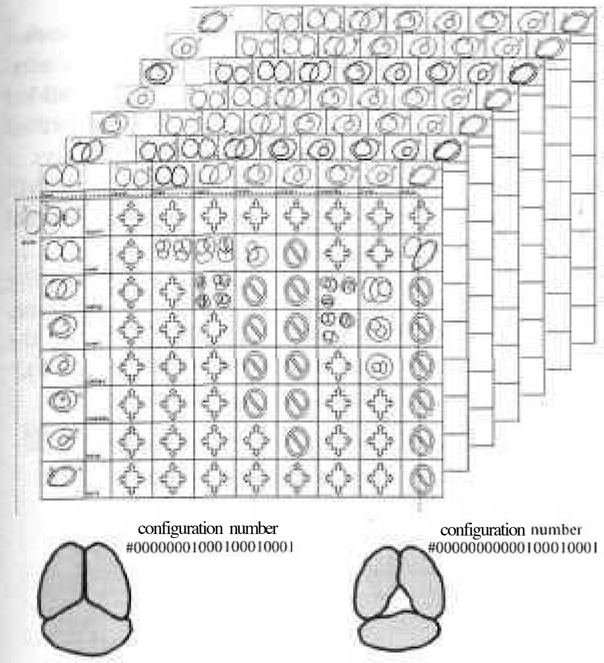

Figure 4: The *20-Intersections* model categorizes 313 possible topological relations with three regions in R_2. Here are two examples of configurations.

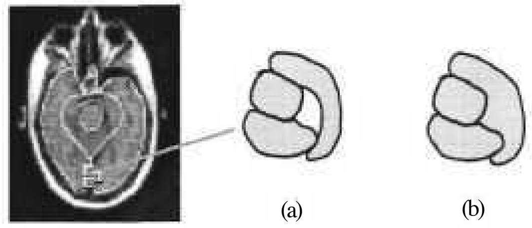

Figure 5: An example of three anomalies in a medical region. In the medical domain, it is crucial that the system can topologically distinguish these two configurations.

Moreover, after modification of spatial relations, it can occur that the integrity of spatial relations can be not respected, e.g.. the configuration <A *Contain* B> ^ <B *Contain* C> ^ <C *Disjoint* A> is impossible. To respect topological integrity, we need to know the spatial expressive power of topological relations with more than two regions. The *20-Intersections* model provides the needed expressive power for three regions.

4.3 Ambiguity resolver

Obviously, certain ambiguities are translated after analysis:

- Physical ambiguity consists in mismatching between semantic objects declared by the user and those of low-level features. MIMS does not currently have such ambiguity because manual physical dimension detection is used.
- Semantic ambiguity occurs when semantic objects or relations proposed by the user are not identified. MIMS currently proposes to differently identify an object or to integrate it into SKB. Whenever semantic relation is defined as a set of spatial ones, ambiguities of semantic relation are resolved at spatial ambiguity level.
- Spatial ambiguities in terms of spatial positioning. As it is known, medical objects location is so complicated to be defined. Next section details how MIMS resolve spatial ambiguities.

The existence of the inference base allows resolving certain ambiguities. The Inference base contains a set of rules used to help in decision-making. This component is upgraded manually but we are working on an automatic upgrade.

5 Spatial ambiguities

The proposed interface in MIMS is hybrid (visual and textual solution), able to give accessibility to a large panel of users. The visual aspect meets important problems of spatial ambiguities. Currently, only few studies have been done to solve ambiguities in visual languages. A taxonomy of ambiguities [18] shows that they are met at different levels and sublevels, making ambiguities one of the most important challenges in visual languages.

A model has been formally defined in [3] for visual systems, which handles ambiguous interpretations of images. This formalism helps us to describe the whole visual environment and to understand where ambiguities can be found. According to the authors, a *visual sentence* is a triple <i,d,<int,mat» where *i* is an *image*, *d* is a *description*, *int* is an *interpretation function* and *mat* a *materialization function*.

Figure 6 shows a representation of the formalism. The *image* is expressed by means of a *pictorial language*. The *description* is expressed by means of a *description language*. The functions *int* and *mat* link *images* with *descriptions*. Several *images* can *materialize* the same *description*. This formalism allows to manage multiple visual representations that convey the same meaning, as needed by different users for different tasks.

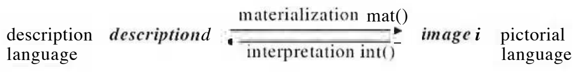

Figure 6: Formalism of a visual language.

We can use this formalism to describe the whole visual environment (Figure 7). The environment encapsulates the system and the user. These two entities communicate through an interface, including pictorial and textual languages. To allow the user to communicate his actions, the interface provides *inputs*. *Inputs* use metaphors, e.g., the user can click on an icon. *Inputs* can also be geometrical shapes, drawn by the user on the screen. The shape drawn can represent the real shape of a spatial object. In MIMS, to visually add or modify a spatial object, the user can directly act on the medical image. For instance, to add an anomaly, he can click on the medical image at the location of the anomaly. He can choose to *materialize* the object with an icon or a generic anomaly shape supplied by the system. The user can also add an object by drawing an object shape on the image. The system checks if the object shape matches a generic anomaly shape.

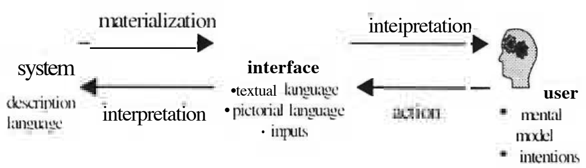

Figure 7: Visual environment.

Figure 7 shows that both user and system must *interpret* the interface. The ambiguity comes from the inteipretation made by the user's viewpoint as well as by the system's viewpoint. Each is capable of making several interpretations. However, which inteipretation is the right one? Ambiguities can be described by the possible answers to the two following questions:

1. How must the system *interpret* user actions?
2. How must the system *materialize* objects and operators to be correctly *interpreted* by the user?

The most important problem with ambiguities is that it induces a lack of precision. It is useless to have a topological model that provides a high topological expressive power if a user description can be interpreted in several ways (several possible configurations in the topological model). For instance, the user *materializes* three anomalies with three icons (Figure 5). He places an icon near the two other icons in the way that all the icons are adjacent. The system can interpret two topological configurations involving the three anomalies.

Several solutions are possible [18]. The first one is that MIMS provides to the user a *clear language,* i.e.. a non-ambiguous textual language. After having visually added or modified spatial objects, the spatial relations derived are displayed in a textual language, in order to give a possibility for the user to check exact semantics. MIMS allows also the user to textually add and modify spatial relations.

Another solution to ambiguities is to establish a *dialog* with the user whenever an ambiguity occurs. For instance, in the previous example, the system shows all the available configurations and requests a choice *(Figure 8).*

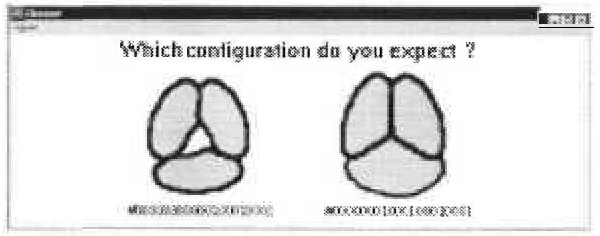

Figure 8: To resolve ambiguities, the system shows all the available configurations and requests a choice.

When checking the respect of the integrity of topological relations, the inference base can find multiple configurations available. The system has to choose one. To solve this ambiguity, the system establishes *dialogs* with the user.

6 Conclusion

The main motivation of this work is to provide a flexible system for answering heterogeneous user queries in medical image databases. Our approach consists in proposing a global image description to achieve an efficient retrieval process able to mix domain and user needs. Ambiguities resolving constitutes a big challenge in medical interfaces where precision is so important in therapeutic treatment and diagnostic analysis. This paper describes our proposition aiming to resolve ambiguities produced by both system and user during storage and query processes. We have implemented a prototype called MIMS (Medical Image Management System). It regroups a set of elements. Spatial Knowledge Model (SKM) is one of the main elements that provides coherent and effective objectivity of image description and analysis. SKM is composed of two components: a Spatial Knowledge Base to assist description and analyzing process, and an Inference Base to check user actions and to manage ambiguities. MIMS is accessible on the WEB and it is currently experimented by several physicians and medicine students. First results and statistics are so satisfactory particularly in terms of high-leveled precision required by the medical domain.

We are currently working on integrating automatic methods able to retrieve low-level features and then to facilitate the user task. Moreover, another study is currently underway which aims to automate SKB structures.

References

[1] Abad-Mota S., Kulikowski C "Semantic Queries on Image Databases: The IMTKAS model". Proc. Of the Basque international Workshop on Information, BIWIT 95, IEEE. Computer Society Press, P. 20-28.

[2] Angelacio, B., Catarci, T.. Santucci, G.. "A Graphical Query Language with Recursion", IEEE Transactions on Software Engineering, Vol 16(10). 1990, P. 1150-1163.

[3] Bottoni. P.. Costabile. M.F. Levialdi, S., Mussio. P.. "Formalising visual languages". In Proc. IEEE Symposium on Visual Languages `95. 1995. P. 45-52.

[4] Ashley J., Flickner M.. Hafner J., Lee D., Niblack W., Petkovic D. "The Query By Image Content System". Proceedings Of the 1995 ACM SIGMOD International Conference on management of data. SIGMOD 95. P. 475.

[5] Aufaure M.-A.. Bonhomme C. LVIS. "A High Level Visual Language for Spatial Data Management", Visual'99. Third International Conference on Visual Information Systems, Amsterdam, The Netherlands, June 2-4, 1999.

[6] Baldi G.. Colombo C. Del Bimbo A.. "Compact and Retrieval-Oriented Video Representation Using Mosaics". Proceedings of Third International Conference, Visual'99, Amsterdam, June 99. P. 171-178.

[7] Bertino E. and Catania B., "A constraint-based approach to shape management in multimedia databases". Multimedia Systems, 1998, P. 2-16.

[8] Bonhomme C, Trépied C. Aufaure M.-A., Laurini R.. "A Visual Language for Querying Spatio-Temporal Databases". ACM GIS99. Kansas City, USA, November 5th-6th, 1999.

[9] Chbeir R., Amghar Y.. A. Flory, "MIMS: A Prototype for medical image retrieval", the 6th international conference of Content Based Multimedia Information Access. RIAO 2000. April 2000, P.846-861.

[10] Chbeir R., Amghar Y., Flory A., "Image Modeling for Medical Databases", ISC A 15th international conference on computers and their applications, CATA 2000, March 2000, P.24-28.

[11] Chbeir R., Amghar Y., Flory A., "System for medical image retrieval: the MIMS model", Third International Conference, Visual'99, Amsterdam, June 99 Proceedings. Page 37-42.

[12] Chetverikov D.. "Pattern regularity as a visual Key", In Proc British Machine Vision Conf.. 1998. P. 23-32.

[13] Wesley W. Chu, Fellow, Chih-Cheng Hsu, Alfonso F. Cárdenas, and Ricky K. Taira. "Knowledge-Based Image Retrieval with Spatial and Temporal Constructs". IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. 6, November/December 1998,

[14] Eakins J.P.. "Automatic image content retrieval: are we going anywhere?", In Proc. Of the 3rd International Conference on Electronic Library and Visual Information Research. May 1996.

[15] Egenhofer, M.. Herring, J., "Categorising Binary Topological relationships between Regions, Lines, and Points in Geographic Databases, in: A Framework for the Definition of Topological Relationships and an Algebraic Approach to Spatial Reasoning within this Framework", Technical Report 91-7, National center for Geographic Information and Analysis, University of Maine, Orono. 1991.

[16] Egenhofer. M.. "Spatial SQL: A Query Presentation Language". IEEE Transactions on Knowledge and Data Engineering, Vol. 6 (1), P. 86-94.

[17] Egenhofer. M.. Query Processing in Spatial Query By Sketch. Journal of Visual Language and Computing. Vol 8 (4). 1997. P. 403-424.

[18] Favetta. F., Aufaure M.-A., "About Ambiguities in Visual GIS Query Languages: a Taxonomy and Solutions", International Conference on Visual Information Systems (Visual'2000). Lyon. France. November 2-4, 2000.

[19] Gelgon M. and Bouthemy P., «A Region Tracking Method with Failure Detection for an Interactive Video Indexing Environment », Third International Conference. Visual'99, Amsterdam. June 99 Proceedings. P. 261-268.

[20] Huet B. and Vailaya A.. " Relational Histograms for shape indexing", IEEE ICCV'98. Jan 1998, P. 563-569.

[21] Ikonomakis N., Plataniotis K.N., Venetsanopoulos A.N., "A Region-based Color Image Segmentation Scheme", Proc. Of the SPIE: Visual Communications and Image Processing, vol. 3653, 1999, P. 1202-1209.

[22] Meyer, B., "Beyond Icons: Towards New Metaphors for Visual Query Languages for Spatial Information Systems", Proceedings of the first International Workshop on Interfaces to Database Systems (R. Cooper ed.), Springer-Verlag, 1993, 113-135.

[23] Shakir, Hussain Sabri, "Context-Sensitive Processing of Semantic Queries in an Image Database System", Information Processing & Management, Vol. 32, No. 5, 1996. P. 573-600.

[24] Zloof, M., M.; "Query-By-Example: A Database Language", IBM Systems Journal, Vol 16(4), 1977, P. 324-343.

[25] Zweigenbaum, "MENELAS: An Access System for Medical Records Using Natural Language", Computer Methods and Programs in Biomedicine, 1994, P. 117-120.

Keyword Association Network: A Statistical Multi-term Indexing Approach for Document Categorization

Kang H. Lee[*] *Judy Kay*[*] *Byeong H. Kang*[+]

[*]Basser Dept of Computer Science
University of Sydney
N.S.W., 2006, Australia
{kangl,judy@cs.usyd.edu.au}

[+]School of Computing
University of Tasmania
Hobart, Tasmania, 7001, Australia
bhkang@utas.edu.au

Abstract

A Keyword Association Network (KAN) is the network of keywords extracted from a collection of documents. In this network, the relationship between keywords is represented by a confidence value. It is argued in this paper that the semantics and importance of a word can be more clearly and accurately measured by making use of other words that are co-occurring in a given document. The term frequency used for measuring the importance of terms in most document categorization methods ignores this important aspect. A KAN is constructed on the basis of co-occurring terms in documents. If two tenns appear more than a certain number of times in the same documents, they are considered as having close relationship. This paper proposes using KAN as a basis for finding informative keywords and using a confidence value in the process of document categorization. The process of constructing and application of KAN for document categorization is presented and the performance comparison with a typical statistical single-term document categorization algorithm - TFIDF classifier - will be shown. The experimental results show that KAN gives significant benefits.

Keywords Document Categorization, Machine Learning, Statistical Multi-term indexing, Semantic-Meaning.

1 Introduction

Term indexing, sometimes known as feature selection or feature extraction, is concerned with extracting informative terms - keywords - from documents based on a weighting scheme. Term indexing has been studied using a growing number of statistical and machine learning techniques and is a crucial part of both document categorization and recommendation systems for textual information. For example, the adjusting parameter (5) extracted terms and their weights are used when calculating the distance or similarity between two documents.

Based on the weighting scheme, term indexing can be broadly grouped into the following three categories: statistical, probabilistic, and information theoretic [7]. In statistical weighting schemes, frequently occurring terms in documents are regarded as informative terms and the term frequency *(TF)* is assigned to term weight. It is also widely recognized that terms which occur in only a few other documents are more informative than ones that appear in many. This consideration results in introducing the inverse document frequency *(IDF)*. These two are combined into a single value, called *TFIDF* [19] and this is widely used for document categorization [8, 11, 12, 13, 22]. Probabilistic weighting schemes are also commonly used for term indexing. When applying such a weighting scheme to document categorization [3, 8, 10], the joint probability of term and category is computed from the training document set and used for the term weights to estimate the probability of a category given a document. Information theoretic methods are based on information gain [15]. In this method, the terms that are concentrated in particular documents are considered as informative terms by measuring signal-noise ratio. In this sense, it operates in a similar way to the *IDF* of statistical weighting schemes.

Term indexing is also categorized into single and multi-term indexing. In single-term indexing on which most previous research has been focused, term weighting schemes are applied to one word without considering the relationship between words. On the other hand, a different point of view on extracting informative terms is that a more accurate and precise meaning of a term can often be identified when looking at other terms in a document. For example, let