Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

PROCEEDINGS OF THE 3rd ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

Large Text Database Visualization

Nathan Combs
TASC
55 Walkers Brook Drive
Reading, MA 01867
ncombs@tasc.com

INTRODUCTION

A system for intuitively visualizing, searching, and querying the contents of large text databases is under development at TASC. This text visualization (TEXTVIZ) system will generate a map-like representation of the contents of a document database which will allow users to visualize how the documents interrelate in terms of their conceptual content. This method of visualizing text databases will allow users to classify documents by the relative similarity of their meaning as well as to discover and to explore conceptual differences between clusters of documents.

TEXTVIZ presents two levels of text database information: a micro and a macro perspective. The micro perspective is focused upon the conceptual content of each document, and the macro perspective graphically links the mosaic of individual micro descriptions into a global picture.

TEXTVIZ uses a topographic map to visualize the contents of a database. The picture of the database presented by this map is global in its scope and is associative and contextual in its nature. The meaning of a document is decided by a limited set of exemplar concepts that have been extracted from its content; these exemplars underlie the macro level classification of the documents.

In contrast, the micro level picture is focused upon the meaning contained within a document. Bridging the macro and micro levels of analysis is a graphical query technique which aids the user to interactively construct a comprehension of the contents of a database. By viewing the distribution of documents on a text map, users can visualize the overall content of the database as well as the content of neighborhoods around specific interest areas.

The TEXTVIZ system which is currently under development involves two major components: a vector-based text processor, and a text map visualization interface. This current development effort extends upon earlier work conducted at TASC in message processing and text database visualization (Carlotto 1992).

VISUALIZING INFORMATION AND LARGE-SCALE TEXT PROCESSING

A text processing system that works with large document databases needs to address two basic challenges: how to extract information about the meaning of documents; and, how to represent this information. It is not enough to produce discriminating descriptions of the documents in the database with a natural language (NL) or text processor. With large databases these document descriptions need to be integrated into a gestalt that is representable and that can be comprehended. There are two separate but related concerns:

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

1. How can the aggregated output of a NL/ text processor be concisely presented?

2. What kind of higher abstraction can be used to express the interrelationships of documents?

One possible method for representing document classifications is by the ranked list. With the ranked list the documents in the database are linearly ordered according to how well they match a target set of concepts. The more relevant a document is to a target set of concepts, the higher in the list it is positioned. How a document relates with other documents with regards to a specific set of criteria can be communicated by this list. One example of a ranked-list system that allowed users to select database objects whose descriptions best fit a user query is given by Rorvig (1991). With this system, if the retrieved objects do not match the query a user can extend the search and look at objects which are "like" the best matching objects found in the list (relevance feedback).

One drawback of the ranked list representation is that the concepts that are being searched for need to be known before the search is implemented. In circumstances where the significant concepts or terminology are not well understood, a ranked linear representation can be restrictive: it does not easily communicate how documents differ from a query, and to what extent. Thus, with Rorvig's example, the results of the relevance feedback are not integrated into a single representation which communicates the relationship of the queries with the contents of the database. Instead, what is presented is a series of disparate "snapshots" of the database as it is evaluated against an evolving query.

Documents can also be represented hierarchically using dendrograms, or trees (a product of single-link clustering, for example). Each leaf in the tree denotes a document. The tree depicts how the documents are incrementally aggregated into ever larger groups: links connect documents or groups of documents to their nearest neighbors. Hierarchical structures are interpreted visually by sequentially traversing their component links: relationships are identified by paths through the cluster hierarchy.

An alternative display for hierarchical document structures that de-emphasizes their sequential interpretation has been proposed by Schneiderman (1991) in his work with tree-maps. With tree-maps, database objects are denoted by surface-filled, color-coded rectangles. Rectangles are colored to show the object type, and the rectangle areas indicate how relevant that object is to its type. While tree-maps are easier to grasp visually than a sprawling tree structure, they, like dendrograms, do not easily communicate how arbitrary documents are related.

The text map is proposed here as an alternative to both the hierarchic structure and the ranked list. This representation provides a comprehensive picture of all documents in a database, unlike the ranked list, and is preferred to hierarchical cluster representations because of its intuitive use of the two dimensional viewing surface.

While hierarchic structures can serve a useful role in facilitating database search and retrieval and thus may underlie the data organization of any representation (van Rijsbergen 1976), for many text database comprehension tasks, hierarchical document displays may be counter-intuitive:

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

1. hierarchies require sequential interpretation.

2. hierarchies restrict comparisons between documents.

3. re-balancing hierarchic structures with new documents can cause dramatic changes to the structure.

A more basic complaint, however, relates to the hierarchic classification methodology itself: because the implicit goal of this method is to partition data into disjoint sets, it cannot easily represent structures derived from statistical distributions (Kohonen 1982). When viewing document distributions, how the parts relate to the distribution carries meaning. It is this connectivity between the documents that is undermined by hierarchic representations: how the parts are connected to the whole is not emphasized, but rather, how the parts can be grouped and re-grouped into ever larger sets.

The text map approach assumes that the interrelationships between documents are significant and representable. It is meant to provide a macro perspective of the database that is intuitive as well as abstract: it will avoid using knowledge about the semantics of the extracted text. The text map interprets documents by their content signature without regard for what the signature actually means. This will allow the system to be used across a variety of applications and databases.

A question then arises. How does a user link specific meaning about the contents of specific documents with a broad intuition about the organization of the database? This is accomplished through an iterative exploration process whereby broad intuitions (the macro picture) about the database are used to guide localized detailed examinations of individual documents (the micro picture) and vice versa. This process will be introduced later in this paper.

Examples of vector-based systems that use visualization maps to represent database information analogous to the approach described here include Carlotto (1992) and Chang (1990).

REPRESENTING TEXT WITH FEATURES

TEXTVIZ will use a NL/ text processing front-end to extract meaning about the contents of a document database; a text map will then be generated that will portray these contents. By comprehending the entire database at once, a user can gain a global perspective of the organization of the contents of the database. By graphically abstracting the output of a NL/ text processor, the relative relationships of the contents of a large database of documents can be characterized. Similarly, clusters of associated documents can be easily identified.

The visualization technique used by TEXTVIZ is designed to be compatible with a range of NL/ text processing paradigms; differing approaches can be selected to reflect differing application requirements. The output of a NL/ text processor may have to be, however, processed by a feature vector translation program so that its output may be cast in a vector format.

Before the output of a NL/ text processor can be graphically portrayed by TEXTVIZ, its output must be translated into a set of describing features. The conceptual content of each document in the

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.1259

database can be represented as a FEATURE VECTOR whose components index specific featu and whose component values measure the degree to which a concept or feature is correlated w a document. What a particular feature "means" conceptually is dependent upon the semantics the processor: it will vary according to the sophistication as well as the domain of the languag model. The features which the system will use to discriminate and classify documents are generated by the NL/ text processor which in turn relies on the customization, training, rules, dictionaries, etc. for its language model.

Consider, for example, a statistical text processor that estimates the meaning of a document by t significant words and phrases that occurs in that document. The significance of a word or phrase measured by the likelihood of that word occurring. If low-likelihood words tend to be more indicative of the meaning of a document than high-likelihood words, then the set of low-frequen words might, for example, be considered as useful describing features for a set of documents. T degree to which a feature correlates with a document can be estimated by the frequency that the associated word or phrase actually occurs within it. By measuring the frequency that each significant word and phrase occurs in a document, a signature or profile of its content can be constructed. This kind of statistical text processor is analogous to the vector-based score-and-ran systems used by Salton (1971) and Stanfill and Kahle (1986).

With more sophisticated NL processors, features may correspond to abstract concepts that are represented by the output symbols and extracted text field values. In these instances, correlation may again be estimated by the frequency that output symbols occur. Figures 1a, 1b, and 1c illustrate a process by which a message can be characterized by a NL processor and then translate into vector form. This approach of representing the content of a document in terms of a vector o features is analogous to vector-based linguistic models of word-sense and semantic discriminatio (eg.Gallant 1991, Miikkulainen and Dyer 1991, McClelland and Kawamoto1986).

ABSTRACTING FEATURES: CLASSIFYING MEANING WITHIN DOCUMENTS

Although the design of the TEXTVIZ system is compatible with a broad range of NL/ and text processing paradigms, the prototype TEXTVIZ system will use an existing in-house text processor The discussion in this section will pertain to this text processor and will focus on how a vector-based text processing methodology can be used to estimate the content of text documents, and how a document signature can be compacted through a process of abstraction.

The developed text processor incorporates four processing stages: a morphological analyzer; a phrase analyzer; a word/phrase vector substituter; and an exemplar selector.

At the first processing stage, the text stream is analyzed morphologically using a set of morphological operators. At the second stage, phrases, expressions are identified and "chunked" using a set of phrase rules. Then at the third stage the recognized words and phrases are translated into their feature vector form and the vectors are aggregated into a document signature. Finally at the fourth stage, exemplar sub-vectors are chosen to represent the document signature vector; the document signature vector is thus abstracted and condensed.

MEANING FROM TEXT
. . . An Example from MUC-3

TST1-MUC3-00

Bogota, 3 April 90 (Inravision Television Cadena 1) – [Report][Jorge Alonso Sierra Valencia][Text] Liberal Senator Federico Estrada Velez was kidnapped on 3 April at the corner of 60th and 48th streets in Western Medellin, only 100 meters from a Metropolitan Police CAI [Immediate Attention Center]. The Antioquia Department liberal party leader had left his house without any bodyguards only minutes earlier. As he waited for the traffic light to change, three heavily armed men forced him to get out of his car and get into a blue renault.

Hours later, through anonymous telephone called to the Metropolitan Police and to the Media, the extraditables claimed responsibility for the kidnapping. In the calls, they announced that they will release the Senator with a new message for the National Government.

Last week, Federico Estrada Velez had rejected talks between the Government and the Drug Traffickers.

From Lehnert and Sundhem, AI Magazine, Fall 1991

TASC

Figure 1a

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592
PROCEEDINGS OF THE 3rd ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP
ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

EXAMPLE (Cont.)

More Sophisticated Case: NL Processor Output

Figure 1b

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

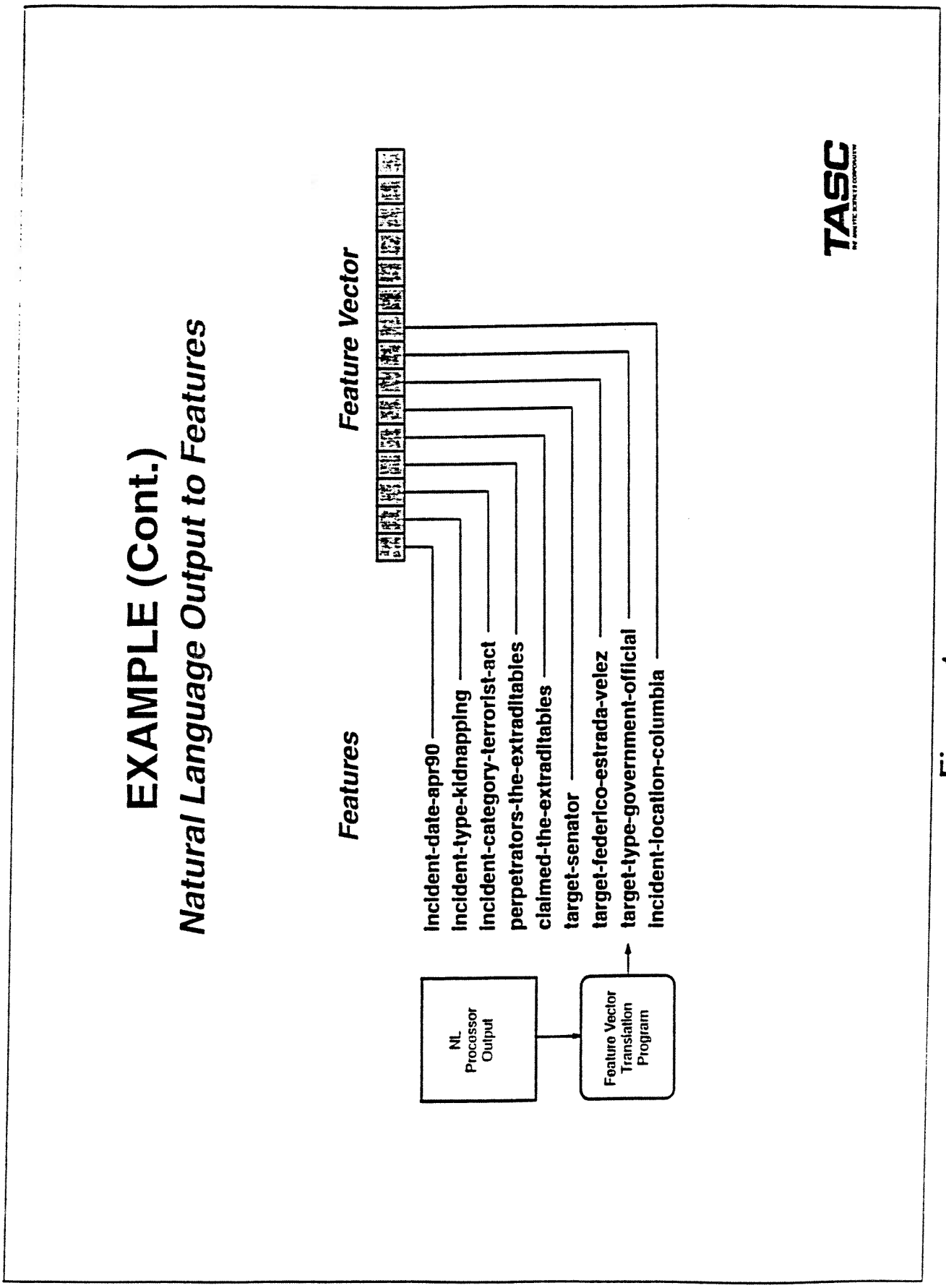

Figure 1c

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

Relevant to the discussion in this paper are the last two processing stages: the word/phrase vector substituter and the exemplar selector. The former exemplifies how a document signature can be constructed and the latter exemplifies how such a signature can be abstracted. Then at the visualization stage, the abstracted signature will be used to graphically classify the document.

The word/phrase vector substituter replaces each word or phrase with a set of feature vectors which signify its meaning. The feature vectors are obtained from a dictionary and are constructed from symbolic definitions. Since a word or phrase may have many definitions and/or synonyms associated with it, each with perhaps its own unique feature vector, how then to select which feature vector to use? In general terms, the approach used here is to look at all possible meanings of all words/phrases in the document and then to cluster them into meaning groups and then to select exemplar definitions (feature vectors) to signify each meaning group. It is assumed that the least likely definitions would disappear from significance as "outliers" on the fringes of the meaning groups.

In this way it is possible to shrink a document signature and to limit the large number of definitions and synonyms that can be linked with the words and phrases in a document. Through this procedure the size of the document signature is reduced by abstracting conflicting or irrelevant information from it.

At the exemplar selection stage, a simulated annealing clustering technique (Kirkpatrick 1983) is used to partition the feature vectors contained by the document signature into N (an arbitrary number) clusters each of which is characterized by an exemplar feature vector. During the clustering process, feature vectors are weighted to reflect their relative importance in determining cluster boundaries: primary definitions are more important than secondary and tertiary definitions, etc.; all definitions are more important than synonyms.

The effect of this clustering procedure is to discard extraneous interpretations of words and phrases as well as to reinforce a consensus in meaning. This consensus is codified by the selection of a set of exemplars which will then represent the document. The manner in which a consensus of meaning is reinforced is analogous to Gallant's 1991 work with developing a vector-based text system that disambiguated semantic meaning in text. Whereas Gallant uses the local context of a word to select which of its possible definitions are most probable, the approach described here uses the global context of all words and their possible interpretations to modulate which of all the possible meanings are most representative. This global approach of estimating meaning is faster to compute though at the cost of being generally less accurate.

TEXT VISUALIZATION: CLASSIFYING DOCUMENTS BY ASSOCIATION

The text map is a visual metaphor that is designed to graphically communicate the taxonomy of the contents of a database.Documents are classified contextually: associations are assumed among proximal documents.

TEXTVIZ converts the information extracted about the content of each document into vectors of numbers; each vector signifies the conceptual content of the associated document: a signature. The macro level representation is calculated from the corpus of vectors. Numeric differences between

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

vectors form the basis of comparison of documents. Thus at the macro level the difference in meaning between documents is "calculated" in terms of the degree that their associated vector patterns differ. The organization of the database is inferred by the system from observing how the patterns of signatures vary across documents: differences between patterns imply differences in meaning.

Similarity between document signatures can be calculated in terms of the distances separating vectors in vector space, or in terms of the angular distances separating vectors. Distance measures do require that the component values be "normalized" across all definitions and is therefore sensitive to the actual magnitudes of the vectors. In contrast, angular distance measures are sensitive to the vector direction or the vector "pattern". This can be advantageous to systems where no standard criteria for measuring correlation between features and their dictionary definitions existed, i.e. measures of correlation where stated empirically, for example. However, this advantage comes at the cost of requiring additional computation.

Ultimately, documents are presented to the user as points on a viewing surface or map; distances between points represent the difference in the estimated meaning of the documents. It is the relative similarity (dissimilarity) of the meaning of documents that is graphically represented by the text map. Thus, there are two levels of abstraction contained by the TEXTVIZ design: first, meaning is abstracted from text using a NL/ text processor; second, the descriptions for all documents are aggregated and then abstracted by a graphical text map display. The abstraction process is accomplished by projecting the document signature vector into a two dimensional visual space (x and y coordinates on the text map).

One method for projecting document signatures onto a two dimensional surfaces uses a Self-Organizing Map (SOM) constructed of a planar network of interconnected processing units (Kohonen1982, 1984). These processing units converge to an "accurate" portrait of the database through an adaptive process based on a competitive neural network learning procedure. Documents are positioned on the map to reflect the relative similarity of their signature vectors. Documents of similar meaning are hence placed close together while documents of dissimilar meaning are placed farther apart. Essentially this process works to project points in vector space (representing documents) into a point on a text map. Hence, similarity reflects the relative distances between vectors in vector space. Other methods for projecting feature vectors onto a graphical viewing surface can also be used (nonlinear mapping: e.g. Sammon 1969; linear mapping: e.g. Friedman et al. 1974).

With very large databases representing each document on the text map may not be practical or desirable. In these cases points on the text map may need to represent a cluster of documents. The actual position of the display point will be calculated using an exemplar of each cluster. A limited hierarchical structure can be constructed "underneath" a text map to provide a more manageable taxonomy of very large databases. This issue will need to be researched before TEXTVIZ is scaled to work with very large document databases.

Figure 2. provides a system overview of the text visualization process. Figure 3. is an example of a text map that was developed by Carlotto (1992). In this map, individual documents are

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

DOCUMENT VISUALIZATION PROCESS

TASC

Process:

- **Features from documents**
- **Features forming concepts**
- **Concepts rendered on a map**

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

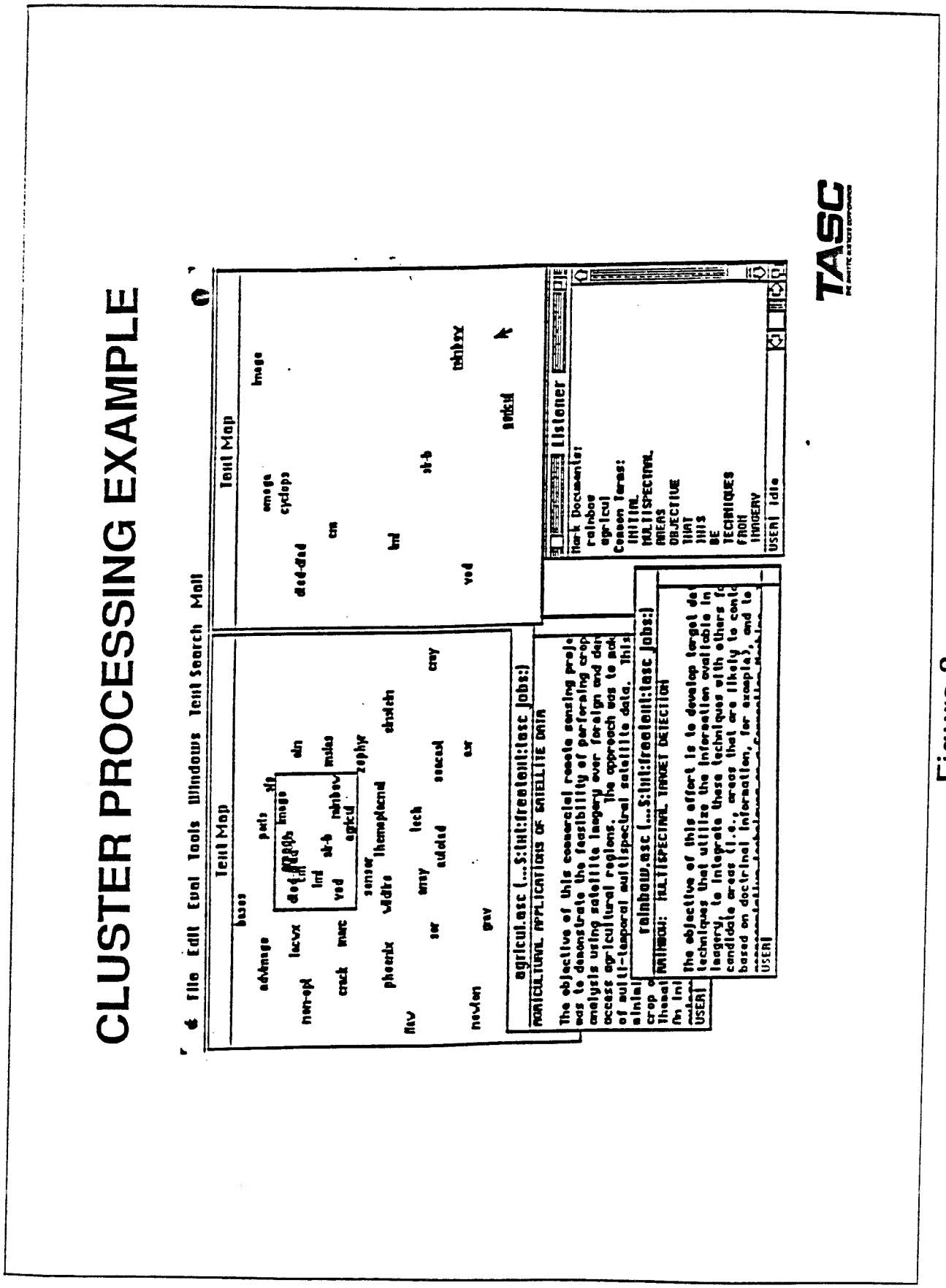

Figure 3

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

represented by word labels. TEXTVIZ will scale this display to work with larger document databases.

QUERYING LARGE TEXT DATABASES

Besides its visualization role, TEXTVIZ will also provide a querying capability that builds upon the intuitive display. To query against the document database, a user defines a region of interest on the graphical visualization map. Documents that are contained by the region can then be either further pruned or all of the documents can be examined in detail. By being able to define queries in terms of sub-spaces or bounded regions on the visualization map, a user can build queries that reflect an understanding of the global distribution of documents as well as knowledge of identified local regions of interest.

With an understanding of the overall distribution of documents in the database, a user can visualize which documents are included and which are not included by a graphical query. This is useful when it comes time to adjust the "scope" of the query. Through a process of selective exploration, the user can sample documents located at the peripheries of the query regions and then adjust the boundaries (query profile) to either exclude or include the documents. Similarly, as whole new clusters of documents of interest are discovered (from the global map), then new query regions can be defined. Finally, just as one can quantify how irrelevant a particular document is to a given query by measuring how far outside a query region it may lie, one can also visualize how good a match a particular document may be by how well it is contained by a query region. These interpretations will be approximate since map resolutions will depend on the quality of the projection and the quality of the NL/ text processing.

The graphical querying technique introduced here offers several advantages over traditional command-line and natural language querying methods. First, it allows the user to conceptualize the contents of the entire document database. This facilitates an understanding of the approximate global scope of the query. Second, users can intuitively appreciate how to adjust their query to include other documents of interest. Queries can be refined to take advantage of additional local information gained about specific regions in the global map. Third, users can define their queries so that they reflect the actual distribution of documents in the database. By choosing to focus upon clusters rather than sparsely populated areas, users can customize their query to the actual contents of the database. Figure 4 outlines this cycle.

As indicated earlier, when working with large databases, the display map may be built on top of a "shallow" hierarchy: points on the map represent clusters of documents rather than individual documents. When querying such a display map, should the query box be abstract or specific? Should the query box, for example, continue to represent the clusters as single points, or should the query box serve as a sort of magnifying glass which "explodes" the clusters and exhibits their full complement? This issue that will need to be researched before the system is scaled to work with very large document databases.

The graphical query cycle is loosely analogous to relevance feedback approaches described by Stanfill and Kahle (1986), Rorvig(1991), and others. Both use empirical information about the

ISSN: 2

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

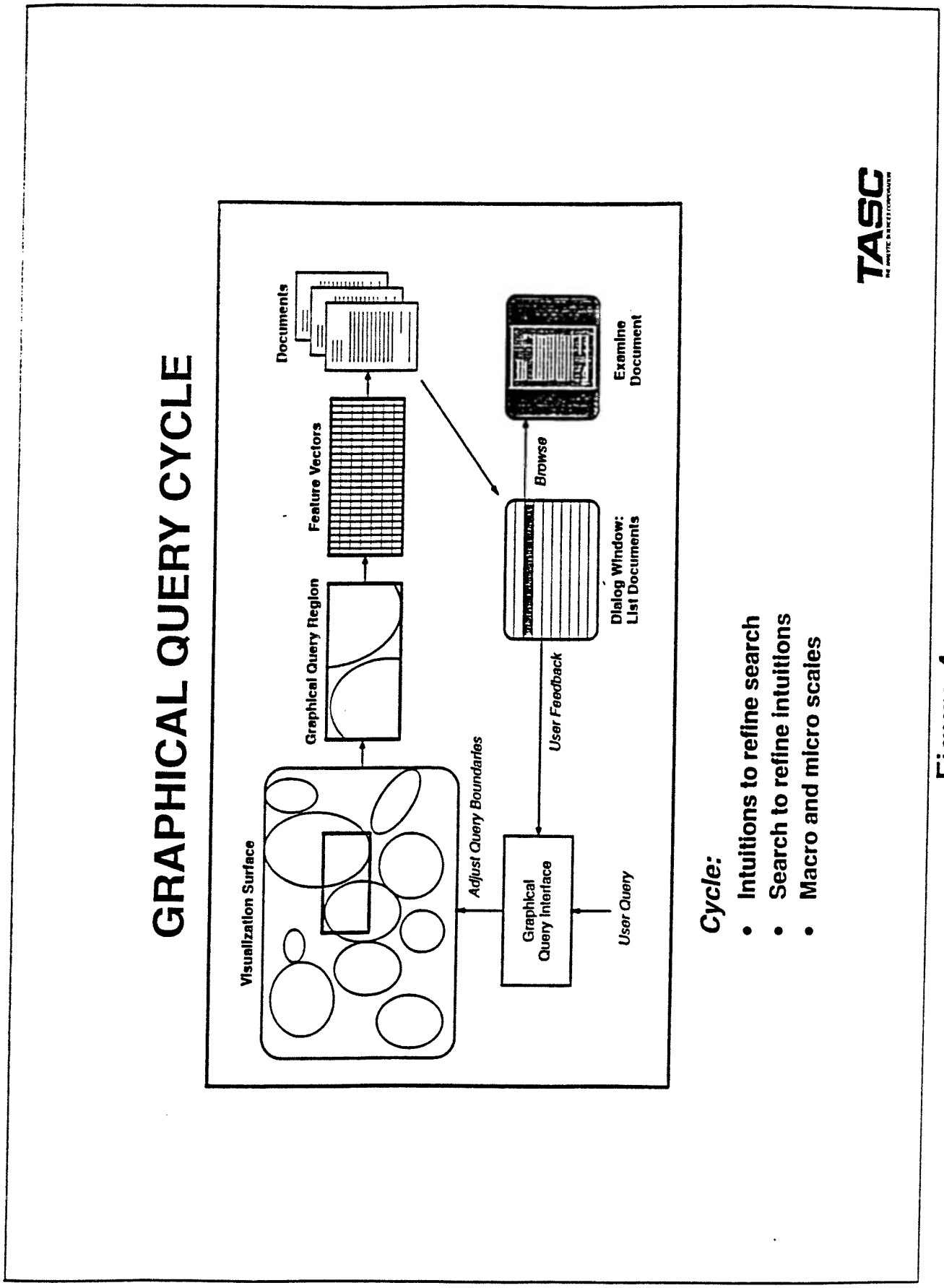

Figure 4

ISSN: 2324-9773

Combs, N. (1992). Large Text Database Visualization. 3rd ASIS SIG/CR Classification Research Workshop, 1-14. doi:10.7152/acro.v3i1.12592

database contents to refine the accuracy of the query. The graphical query method, however, is based on a graphical metaphor and a text map abstraction.

DISCUSSION

TEXTVIZ is designed to provide an intuitive means of viewing the contents of large document databases. The text map is a graphical abstraction which allows documents to be classified by the relationship of their content to the rest of the database. The idea that document taxonomies can be defined graphically by regions on the text map is implicit to the described graphical query method.

TASC's first test of the TEXTVIZ approach will be to visualize and analyze the TIPSTER database of text articles; results will be reported to the NIST/DARPA Text Retrieval Conference (TREC) in November 1992.

WORKS CITED

Carlotto, M. "A Text-Based Geographic Information System." *TASC white paper.* 1992.

Chang, S. "Visual Reasoning for Information Retrieval from Very Large Databases." *Journal of Visual Languages and Computing*, 1. 1990.

Friedman, J., and J. Tukey. "A Projection Pursuit Algorithm for Exploratory Data Analysis." *IEEE Transactions on Computers*, Vol.C-23, No. 9, Sept. 1974.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. *Science*, vol 220. 1983.

Kohonen T. *Self-organization and associative memory.* Berlin: Springer-Verlag. 1984.

Kohonen T. "Clustering, Taxonomy, and Topological Maps of Patterns." *Proceedings of the 6th ICPR*, Oct 19-22, 1982.

McClelland, J.L., and Kawamoto, A.H. "Mechanisms of sentence processing: Assigning roles to constituents." In J.L. McClelland, and D.E. Rumelhart (eds). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations.* Cambridge, MA: MIT Press. 1986.

Miikkulainen, R., and M. Dyer. "Natural Language Processing with Modular PDP Networks and Distributed Lexicon." *Cognitive Science* 15, 1991.

Rijsbergen, C.J. *Information Retrieval.* Butterworths: Boston. 1979.

Rorvig, M. E. "A Vector-Product Information Retrieval System Adapted to Heterogeneous, Distributed, Computing Environments." *Technology 2001, NASA conference*, Dec. 3-5, 1991.

Salton, G. *The SMART Retrieval System - Experiment in Automatic Document Processing.* Prentice-Hall: Englewood Cliffs, NJ. 1971.

Sammon, J. "A nonlinear mapping algorithm for data structure analysis" *IEEE Transactions on Computers.* Vol C-18, No. 5, May1969.

Shneiderman, B. "Visual User Interfaces for Information Exploration." *Department of Computer Sciences Technical Report No. 2748.* August 1991.

Stanfill, C., and Kahle, B. "Parallel Free-Text Search on the Connection Machine System." *Communications of the ACM*, Vol 29, 12. December 1986.

ISSN: 2324-9773