

Text Clustering: Algorithms, Semantics and Systems

Joshua Zhexue Huang¹ & Michael Ng,²
& Liping Jing¹

¹ The University of Hong Kong
² Hong Kong Baptist University

April 9, 2006
PAKDD06 Tutorial
Singapore

Section 0

Introduction

Introduction to Text Clustering

- Text data is ubiquitous.
- As the volume of text data increases, management and analysis of text data becomes unprecedentedly important.
- Text mining is an emerging technology for handling the increasing text data.
- Text clustering is one of the fundamental functions in text mining.
 - Text clustering is to divide a collection of text documents into different category groups so that documents in the same category group describe the same topic, such as classic music or Chinese history.

8/3/2006

PAKDD2006

3

Challenges

- Unlike clustering structured data, clustering text data faces a number of new challenges.
 - Volume,
 - Dimensionality,
 - Sparsity, and
 - Complex semantics.
- These characteristics require clustering techniques to be scalable to large and high dimensional data, and able to handle sparsity and semantics.

8/3/2006

PAKDD2006

4

Objectives of This Tutorial

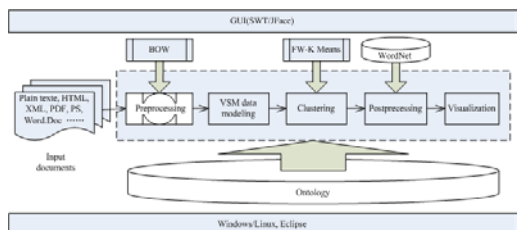
- Introduce techniques of text clustering, including
 - Text data representation techniques and preprocessing methods that are used to convert original text in various formats into a representation model.
 - Use of ontology to enhance semantic representation of the original model.
 - Classical clustering algorithms that have been used on text data, in particular, the recently developed k-means type subspace clustering algorithms and their applications to text data.
 - Techniques to use ontology to extract representative terms and concepts to represent the clustering results and some visualization methods.
 - Some existing text clustering systems and applications.

8/3/2006

PAKDD2006

5

Text Clustering Architecture



8/3/2006

PAKDD2006

6

Outline

1. Text data and representation models
2. Text data preprocessing techniques
3. Ontology and semantic enhancement of presentation models
4. Text clustering algorithms
5. Post-processing techniques with ontology and clustering visualization
6. Text clustering systems and applications

Section 1

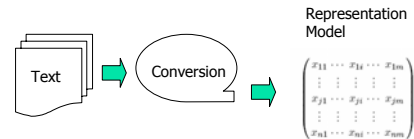
Text Data and Representation Models

Text Data

- Text data occur in different formats
 - Plain text
 - DOC
 - PDF
 - PS
 - HTML
 - XML
 - Email
 - ...
- Different representation formats offer different capabilities to describe context, structure, semantics, presentations of text content

Representation Model

- In information retrieval and text mining, text data of different formats is represented in a common representation model, e.g., Vector Space Model
- Text data is converted to the model representation



Different Representation Models

- Probabilistic model
- Vector space model --- VSM
- Ontology-based VSM

Probabilistic Model

- Description
 - Documents are represented by means of a probability distribution of terms
$$P(t_1, \dots, t_m)$$
 - This model specifies the probability $P(d_j | L_c)$ that the document d_j belongs to the category C
- Advantage --- documents are ranked according to their probability of being relevant to a category
- Disadvantage --- it needs to guess the initial separation of documents into relevant and non-relevant sets, and the terms are considered to be independent

From S.E. Robertson and K.S. Jones, Relevance weighting of search terms. *Journal of the American society for Information Sciences*, 27(3): 129-146, 1976.

Vector Space Model (VSM)

- The most popular representation model used in information retrieval and text mining.
- In VSM, a text document is represented as a vector of terms $\langle t_1, t_2, \dots, t_p, \dots, t_n \rangle$.
- Each term t_i represents a word or a phrase.
- The set of all n unique terms in a set of text documents forms the vocabulary for the set of documents.
- A set of documents are represented as a set of vectors, that can be written as a matrix.

8/3/2006

PAKDD2006

13

Matrix Representation of VSM

- A document collection is represented as a matrix:

$$\begin{pmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jm} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{ni} & \cdots & x_{nm} \end{pmatrix}$$

- where each row represents a document, each column indicates a term, and each element x_{ji} represents the frequency of the i^{th} term in the j^{th} document.

8/3/2006

PAKDD2006

14

Representation of Terms

- Three ways to represent a term value x_{ji}
 - Frequency representation:
 - x_{ji} is the frequency of term i in document j
 - Binary representation
 - $x_{ji} = 1$ indicates that term i occurs in document j , otherwise, $x_{ji} = 0$
 - Term frequency-inverted document frequency (*tfidf*)

$$tfidf(d_j, t_i) = tf(d_j, t_i) \times \log\left(\frac{|D|}{df(t_i)}\right)$$

- Where $tf(d_j, t_i)$ is the frequency of term t_i in document d_j , $|D|$ is the total number of documents, and $df(t_i)$ is the number of documents in which t_i occurs.

8/3/2006

PAKDD2006

15

Advantages and Disadvantages of VSM

- Advantages
 - Simple
 - Easy to calculate similarity between two documents
 - Data mining algorithms can be applied directly to text data
- Disadvantages
 - Terms are assumed independent (which is not true in the real text document)
 - Lack of semantics
 - High dimensionality and sparsity

8/3/2006

PAKDD2006

16

Ontology-based VSM

- Consider the relationship between terms
- Introduce semantic concepts into data models
- Combine ontology with the traditional VSM
- Terms (dimensions) have semantic relationships, rather than independent

8/3/2006

PAKDD2006

17

Important References

- W. Fan, L. Wallace, S. Rich, and Z. Zhang, Tapping into the power of text mining, the Communications of ACM, 2005.
- M.M. Gomez, A.L. Lopez, and A.F. Gelbukh, Information retrieval with conceptual graph matching, In DEXA, 312-321, 2000
- A. Hotho, S. Staab, and G. Stumme, Text Clustering based on background knowledge, TR425, AIFB, German, 2003
- J.M. Ponte and W.B. Croft, A language modeling approach to information retrieval, In research and development in information retrieval, 275-281, 1998
- S.M. Weiss, N. Indurkha, T. Zhang, and F.J. Damerau, Text mining, Springer, 2005
- R. Yate and B. Neto, Modern information retrieval, Addison Wesley, 1999
- W. Yang, J.Z. Huang and M.K. Ng, A data cube model for prediction-based web prefetching, Journal of intelligent information systems, 20(1), 11-30, 2003

8/3/2006

PAKDD2006

18

Section 2

Text Preprocessing Techniques

Text Preprocessing Techniques

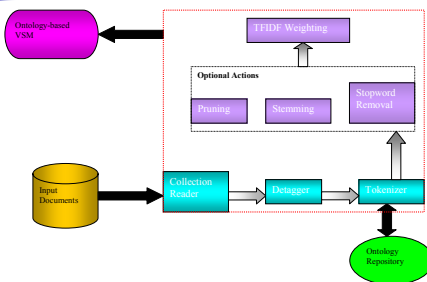
- Objective
 - Transform unstructured or semi-structured data (text data) into structured data model (VSM) or ontology-based VSM

8/3/2006

PAKDD2006

20

Preprocessing Operations and Process



8/3/2006

PAKDD2006

21

Techniques for Preprocessing

- Collection reader
- Detagger
- Tokenizer
- Stopword removal
- Stemming
- Pruning
- Term weighting

8/3/2006

PAKDD2006

22

Collection Reader

- Transform raw document collection into a common format, e.g., XML
- Use tags to mark off sections of each document, such as, <TOPIC>, <TITLE>, <ABSTRACT>, <BODY>
- Extract useful sections easily
- Example
 - “Instead of direct prediction of a continuous output variable, the method discretizes the variable by kMeans clustering and solves the resultant classification problem.”

8/3/2006

PAKDD2006

23

Detagger

- Find the special tags in document
 - “ “ ” ” ”
 - ” ” ”
- Filter away tags
 - “Instead of direct prediction of a continuous output variable the method discretizes the variable by kMeans clustering and solves the resultant classification problem”

8/3/2006

PAKDD2006

24

Tokenizer

- Define tokens
 - Nonempty sequence of characters, excluding spaces and punctuations, e.g., “ ”
- Represent token
 - A suitable table (indexing token positions and the id of the document in which tokens occur)
- Optional functions
 - Removing stopwords
 - Stemming
 - Pruning

8/3/2006

PAKDD2006

25

Removing Stopwords

- Stopwords
 - Function words and connectives
 - Appear in a large number of documents and have little use in describing the characteristics of documents
- Remove stopwords
 - Don't need to index stopwords
 - Reducing index space and improving performance
 - Issues
 - Queries containing only stopwords ruled out
 - Polysemous words that are stopwords in one sense but not in others (E.g.; *can* as a verb vs. *can* as a noun)

8/3/2006

PAKDD2006

26

Example

- Removing Stopwords
 - Stopwords:
 - “of”, “a”, “by”, “and”, “the”, “instead”
 - Example
 - “Instead direct prediction of a continuous output variable the method discretizes the variable by kMeans clustering and solves the resultant classification problem”

8/3/2006

PAKDD2006

27

Stemming

- Conflate words to help match a term with a morphological variant in the corpus
- Remove inflections that convey parts of speech, tense and number
 - E.g.: *University* and *Universal* both stem to *Universe*
- Techniques
 - Morphological analysis (e.g., Porter's algorithm)
 - Dictionary lookup (e.g., WordNet)
- Increase recall at the price of precision
 - Abbreviations, polysemy and names coined in the technical and commercial sectors
 - E.g.: Stemming “ides” to “IDE”, “SOCKS” to “sock” may be bad!

8/3/2006

PAKDD2006

28

Example

- Stemming
 - Stems:
 - “prediction ---> predict”
 - “discretizes ---> discretize”
 - “kMeans ---> kMean”
 - “clustering --> cluster”
 - “solves ---> solve”
 - “classification ---> classify”
 - Example sentence
 - “direct predict continuous output variable method discretize variable kMean cluster solve resultant classify problem”

8/3/2006

PAKDD2006

29

Pruning

- Discard words appearing rarely or more frequently
- Rationale
 - Infrequent or more frequent terms do not help with identifying appropriate clusters rather than adding noise to the distance measures and degrading the overall performance
- Techniques
 - Term frequency (δ_1, δ_2 are pre-defined thresholds)

$$\sum_{d \in D} tf(d, t) \geq \delta_2 \quad \text{or} \quad \sum_{d \in D} tf(d, t) \leq \delta_1$$

- Document frequency (β_1, β_2 are pre-defined thresholds)

$$\sum_{d \in D} \varphi(tf(d, t)) \geq \beta_2 \quad \text{or} \quad \sum_{d \in D} \varphi(tf(d, t)) \leq \beta_1$$

8/3/2006

PAKDD2006

30

Weighting Terms

- Weight the frequency of a term in a document
- *tfidf*

$$tfidf(d, t) := tf(d, t) \times \log\left(\frac{|D|}{df(t)}\right)$$

- Rationale
 - Not all terms are equally useful
 - Terms that appear too rarely or too frequently are ranked lower than terms that balance between the two extremes
 - Higher weight means that the term is better to contribute to clustering results

Important References

- McCallum, A.K., Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, <http://www.cs.cmu.edu/mccallum/bow>, 1996.
- G. Salton, Automatic text processing: the transformation, analysis and retrieval of information by computer, Addison-Wesley, 1989
- <http://www.lextek.com/manuals/onix/stopwords1.html>
- M.F. Porter, An algorithm for suffix stripping, Program, 14(3), 130-137, 1980
- G. Amati, C. Carpineto, and G. Romano, Pub at trec-10 web track: A probabilistic framework for topic relevance term weighting In the tenth text retrieval conference, 2001
- N. Ide and J. Veronis, Introduction to the special issue on word sense disambiguation: the state of the art, Computational Linguistics, 24(1), 1-40, 1998

Section 3

Ontology and Semantic Enhancement of Presentation Models

Ontology and Semantic Enhancement of Presentation Models

- Why ontology-based model?
- How to represent ontology?
- How to compile ontology into text presentation models (VSM)?

Why Ontology-based Model?

- Provide a high level structural view for navigation through mostly unknown terrain
- Represent unstructured data (text documents) according to ontology repository
 - Each term in a vector is a concept rather than only a word or phrase
- Determine the similarity of documents
 - Based on the distance of document term and concept vectors

How to Represent Ontology for Text Data?

- Two typical methods
 - Paradigm: Specific concepts are represented in
 - Time: interval logic or point representation
 - Hierarchy: different levels
 - Concept description:
 - the way in which a hierarchy is built
 - publication<scientific publication<book<dissemination
 - publication<book<scientific book<dissemination
 - distinctions are made by attributes or subclasses

Methods to Represent Ontology

- Conceptualization ontology
 - **Model coverage and granularity:** which things are described and in how much detail
 - One ontology only models cars, not trucks
 - One models trucks, but only with very light hierarchy
 - One models trucks with fine-grained (deep) hierarchy
 - Two ontologies both model trucks in detail, but to different user communities (truck producer vs. truck repairer)
 - **Scope of concept:** what are the instances of a concept
 - employee can mean “all people that have a room within a company” or “all people that get paid by a company”

8/3/2006

PAKDD2006

37

Methods to Represent Ontology

- Terminological ontology
 - **Synonyms:** several words for the same concept
 - employee (HR)=staff (Administration)=researcher (R&D)
 - car=automobile
 - **Homonyms:** one word with several meanings
 - bank: river bank vs. financial bank
 - fan: cooling system vs. sports fan
 - **Encoding / Language:** what linguistic or metric system is used
 - English or Dutch
 - inches or centimeters

8/3/2006

PAKDD2006

38

How to Compile Ontology into VSM?

- Two approaches
 - **Hotho et al., 2003 “Text clustering based on background knowledge”**
 - Strategies for concepts
 - Strategies for disambiguation
 - Strategies for hypernyms
 - **Jing et al., 2006 “Ontology-based distance measure for text clustering”**
 - Semantic term similarity

8/3/2006

PAKDD2006

39

Ontology-based VSM (Hotho et al.)

- A document vector with ontology consideration is represented by:

$$t_d := \langle tw(d, t_1), \dots, tw(d, t_m), cw(d, c_1) \dots, cw(d, c_l) \rangle$$

where “ tw ” is the frequency of term t_i in document d , “ cw ” is the frequency of concept c_j in document d

- Three methods to calculate the concept frequency.

8/3/2006

PAKDD2006

40

Strategies for Concepts

- Strategies for concepts
 - Add concepts (“add”)
 - Extend term vector t_d by new entries for concepts c_d appearing in the document set
 - A term that appears in c_d would be accounted at least twice in the new vector
 - Replace terms by concepts (“repl”)
 - Expel all terms from t_d for which at least one corresponding concept exists
 - Concept vector only (“only”)
 - Discard terms that do not appear in concept repository; only c_d represents the document vector

8/3/2006

PAKDD2006

41

Strategies for Disambiguation

- Strategies for disambiguation (in “add” and “repl” concept strategies)

- All concepts (“all”)
 - Consider all concepts for augmenting the text document representation
- First concepts (“first”)
 - Consider only the first appearing concept in the ordered list of concepts
- Disambiguation by context (“context”)
 - Consider the semantic vicinity of the concept and select the concept which has maximum frequency

8/3/2006

PAKDD2006

42

Strategies for Hypernyms

- Strategies for hypernyms
 - For some concept, there may be sub-concepts in the taxonomy
 - Like:

$$H(c, r) = (c' \mid \exists c_1, \dots, c_i \in C : c' \prec c_1 \prec \dots \prec c_i = c, 0 \leq i \leq r)$$
 - $r=0$
 - The strategy does not change the given concept frequency
 - $r=k$
 - The strategy adds to each concept the frequency counts of all sub-concepts in the k levels below it in the ontology
 - $r=\infty$
 - The strategy adds to each concept the frequency counts of all its sub-concepts

8/3/2006

PAKDD2006

43

Ontology-based VSM (Jing et al. 2006)

- Each element of a document vector considering ontology is represented by:

$$\tilde{x}_{j_i} = x_{j_i} + \sum_{\substack{i_2=1 \\ i_2 \neq i_1}}^m \delta_{i_1 i_2} x_{j_{i_2}}$$

where x_{j_i} is the original frequency of term t_i in the j^{th} document, $\delta_{i_1 i_2}$ is the semantic similarity between term t_{i_1} and term t_{i_2} .

8/3/2006

PAKDD2006

44

Example

- According to WordNet, terms 'ball', 'football', and 'basketball' are semantically related to each other. Updating document vectors in Table 1 by the formula, new ontology-based vectors are obtained.

Table 1: A simple example for traditional *term-based VSM*

	ball	football	basketball	food
d1	5	0	3	2
d2	0	4	1	0

Table 2: The representation for Table 1 data in *ontology-based VSM*

	ball	football	basketball	food
d1	7.4	6.4	7	2
d2	4	4.8	4.2	0

8/3/2006

PAKDD2006

45

Semantic Similarity between Two Terms: Method 1 (constructing ontology)

- Parse a given large corpora and extract the syntactic dependencies (object/attribute pair), such as, house_subj(museum), combine_obj(abstraction), allude_to(influence)
- Weight object/attribute pairs with existing information measures.
- Calculate the mutual term similarity with Cosine, Jaccard, L₁ norm, etc. Here, two terms are considered to be mutually similar if one occurrence of an attribute with one term is also counted as an occurrence of that attribute with other term.

From Ciminao et al., 2005 "learning concept hierarchies from text corpora using formal concept analysis"

8/3/2006

PAKDD2006

46

Semantic Similarity between Two Terms: Method 2

- Directly assign the value with WordNet.
 - If $t_{i_1} \in \text{Synsets}(t_{i_2})$ or $t_{i_2} \in \text{Synsets}(t_{i_1})$, then $\delta_{i_1 i_2}$ is set to be a pre-defined value δ .

- Based on the new ontology-based VSM, calculate the mutual information matrix M .
 - Orthogonalize $M=BB^T$ to get its factor B .

- Update VSM with the correlation factor matrix B .

$$\hat{X}_j = X_j B$$

- Where X_j is the j^{th} document vector in the original VSM

From Jing et al., 2006 "Ontology-based distance measure for text clustering"

8/3/2006

PAKDD2006

47

Term Similarity Calculated by Method 2

- According to WordNet, terms 'ball', 'football', and 'basketball' are semantically related to each other. Updating document vectors in Table 1 by the formula, new ontology-based vectors are obtained.

(t_1, t_2)	Term Similarity
(software, hardware)	0.9240
(Arab, people)	0.9163
(baseball, sport)	0.8974
(space, science)	0.8948
(graphics, compute)	0.8365
(Jew, race)	0.7769
(orbit, satellite)	0.7514
(symmetric, circle)	0.7212
(team, player)	0.7028
(science, research)	0.6113

8/3/2006

PAKDD2006

48

Experiments on Real Dataset

- Text data *four subsets of 20-News* groups

Categories	A4(n_d)	A4U(n_d)
comp.graphics	100	120
rec.sport.baseball	100	100
sci.space	100	59
talk.politics.mideast	100	20
Categories	B4(n_d)	B4U(n_d)
comp.graphics	100	120
comp.os.ms-windows	100	100
rec.autos	100	59
sci.electronics	100	20

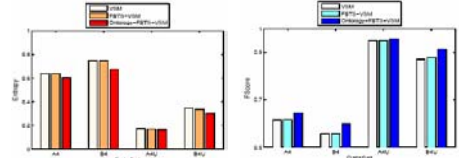
8/3/2006

PAKDD2006

49

Experimental Results of Text Clustering with Ontology-based VSM

- Comparison of clustering quality (Entropy & FScore) for *standard k-means* using different text representation



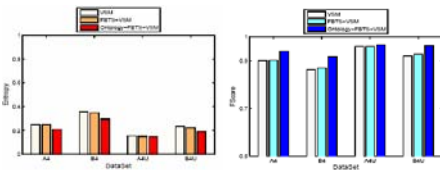
8/3/2006

PAKDD2006

50

Experimental Results

- Comparison of clustering quality (Entropy & FScore) for *feature weighting k-means* using different text representation



8/3/2006

PAKDD2006

51

Experimental Results of Text Clustering with Ontology-based VSM

- Relative improvements of clustering quality (Entropy (RIEn) & Fscore (RIFS)) for standard k-means and feature weighting k-means

Methods	A4			
	B4	A4U	B4U	
<i>PW-KMeans</i>	RIFS(%) 4.38	6.18	0.69	4.61
	RIEn(%) 16.02	17.24	4.75	18.37
<i>standard k-means</i>	RIFS(%) 4.80	7.35	0.88	4.91
	RIEn(%) 5.71	9.89	4.10	13.12

- RIEn and RIFS achieved on the datasets with similar topics (B4 and B4U) are generally higher than those achieved on the datasets with different topics (A4 and A4U).

8/3/2006

PAKDD2006

52

Future Work

- Comparison of different methods of calculating term similarity that will affect the text clustering quality.
- Construct proper data model to store the large volume text corpora, which takes into account the semantic information.
- Apply the term similarity and data model to the other text mining techniques, e.g., text classification.

8/3/2006

PAKDD2006

53

Important References

- S. Bloehdorn, P. Cimiano, A. Hotho, and S. Staab, An Ontology-based framework for text mining, LDV Forum – GLDV Journal for computational linguistics and language technology, 20(1), 87-112, 2005
- C. Fellbaum, *Wordnet: an electronic lexical database*, the MIT press, 1998
- M. Gruninger and J. Lee, Ontology applications and design, Communication of the ACM, 45(2), 39-41, 2002
- A. Hotho, S. Staab, and G. Stumme, Text Clustering based on background knowledge, TR425, AIFB, German, 2003
- A. Hotho, S. Staab, and G. Stumme, WordNet improves text document clustering, In Proc. of the SIGIR on Semantic Web Workshop, 2003
- L. Jing, L. Zhou, M.K. Ng, and J.Z. Huang, Ontology-based distance measure for text clustering, In Proc. Of the SIAM SDM on Text Mining Workshop, 2006

8/3/2006

PAKDD2006

54

Section 4

Text Clustering Algorithms

Text Clustering Algorithms

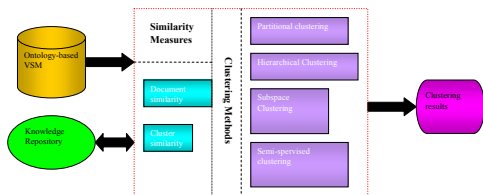
- Objective
 - Efficiently and automatically grouping documents with similar content into the same cluster
- Core
 - Similarity measures
 - Clustering methods

8/3/2006

PAKDD2006

56

Block Diagram



8/3/2006

PAKDD2006

57

Document Similarity Measures

- There are many different ways to measure how similar two documents are, or how similar a document is to a query
- Highly depending on the choice of terms to represent text documents
 - Euclidian distance (L_2 norm)
 - L_1 norm
 - Cosine similarity

8/3/2006

PAKDD2006

58

Document Similarity Measures (Cont'd)

- Euclidian distance (L_2 norm)
$$L_2(\vec{x}, \vec{y}) = \sum_{i=1}^m (x_i - y_i)^2$$
- L_1 norm
$$L_1(\vec{x}, \vec{y}) = \sum_{i=1}^m |x_i - y_i|$$
- Where \vec{x} and \vec{y} represent one text document respectively, x_i and y_i are the i th term corresponding to vector \vec{x} and \vec{y}

8/3/2006

PAKDD2006

59

Document Similarity Measures (Cont'd)

- Cosine similarity
 - For two vectors \vec{x} and \vec{y} , the cosine similarity between them is given by:
$$\cos(\angle(\vec{x}, \vec{y})) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$
 - Here $\vec{x} \cdot \vec{y}$ is the vector product of \vec{x} and \vec{y} , calculated by multiplying corresponding frequencies together
 - The cosine measure calculates the angle between the vectors in a high-dimensional data space

8/3/2006

PAKDD2006

60

Cluster Similarity Measures

- Computing similarity / coherence of two clusters:
 - **Single linkage**: similarity of two most similar document vectors counts
 - **Complete linkage**: similarity of two least similar document vectors counts
 - **Group-Average**: average similarity of all document vectors is calculated

8/3/2006

PAKDD2006

61

Clustering Methods

- General clustering methods
- Clustering methods for text data

8/3/2006

PAKDD2006

62

General Clustering Methods

- Partition based Clustering:
 - User specified k representative points taken as cluster centers and points assigned to cluster centers
 - k -means, k -medioids, CLARANS (VLDB 94), BIRCH (SIGMOD 96),...
- Hierarchical Clustering:
 - Each point is a cluster.
 - Merge *similar* points together gradually.
 - CURE (SIGMOD 98)

8/3/2006

PAKDD2006

63

General Clustering Methods

- Categorical Clustering:
 - Clustering of categorical data e.g automobile sales data: color, year, model, price, etc
 - Best suited for non-numerical data
 - concepts
 - CACTUS (KDD 99), STIRR (VLDB 98)

8/3/2006

PAKDD2006

64

What's Wrong with Them?

- Curse of dimensionality
 - Ten or hundred thousand dimensions in text representation (e.g., 19949 documents --- 43586 words in 20-NewsGroup)
 - The distance between every pair of documents in high dimensions is almost the same for a variety of distance functions

8/3/2006

PAKDD2006

65

What's Wrong with Them? (Cont'd)

- Validate the cluster
 - Traditionally clustering methods fail to evaluate the clustering results
 - Statistical tests in high dimensions/baseline distributions
 - Fail to interpret the clustering results
 - *k*-means only gives the mean frequency or weight of the terms in each cluster, but cannot say which one is keyword because higher frequency or weight does not represent the term is important

8/3/2006

PAKDD2006

66

Text Clustering Methods

- Subspace Clustering
- Text data: large/high dimensions/sparse
- Low-dimensional clusters embedded in the data space
- Clusters contain in different subspaces
- Dimensions in each subspace may be different
- Different from feature selection

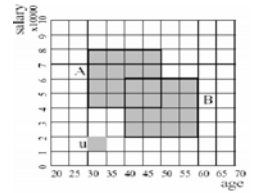
8/3/2006

PAKDD2006

67

Why Subspace Clustering?

- Background
 - Some document groups are correlated with a given set of terms and others are correlated with respect to different terms.
 - As shown in the right figure, cluster A and B are correlated with different set of terms



8/3/2006

PAKDD2006

68

Subspace Clustering

- Define the subset of terms that are relevant to each cluster
 - Subspaces are allowed to be overlapping
- The cluster found should represent some meaningful pattern in the data in the context of the particular domain
- Scale well with respect to the number of documents and the number of dimensions in large datasets
- Also scale with respect to the number of dimensionality of the subspace where the clusters are found

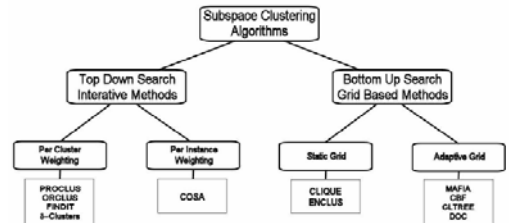
8/3/2006

PAKDD2006

69

Subspace Clustering (Cont'd)

- Two major types of subspace clustering algorithms distinguished by subspace searching methods



8/3/2006

PAKDD2006

70

Bottom-Up Subspace Clustering

- CLIQUE (SIGMOD 98)
- ENCLUS (SIGKDD99)
- MAFLA (CPDC-TR99)
- CBF (ACM Press02)
- CLTREE (ACM Press00)
- DOC (SIGMOD02)

8/3/2006

PAKDD2006

71

CLIQUE

- CLIQUE (SIGMOD 98) – automatic subspace clustering of high dimensional data for data mining application
 - Identification of dense units
 - Identification of clusters
 - Generation of minimal description
 - MDL principle are used as pruning method
 - Encode the input data under a given model and select the encoding that minimizes the code length

8/3/2006

PAKDD2006

72

CLIQUE (Cont'd)

- CLIQUE (SIGMOD 98) – User specifies the grid size and threshold for each dimension
 - Finer grids: enormous computation and coarser grids : loss of quality
 - Noise is another consideration in Finer grids
 - A bottom-up algorithm by combining *dense* regions in different subspaces
 - A hyper-rectangle in a multidimensional space is *dense* if it contains more points than a *user specified* threshold percentage of the total number of points
 - CDUs (candidate dense unit) in any dimension k is formed by combining dense units of dimension $(k-1)$ which share the first $(k-2)$ dimensions

8/3/2006

PAKDD2006

73

Others

- ENCLUS (SIGKDD99) – Entropy-based subspace clustering for mining numerical data
- MAFIA (CPDC-TR99) – Efficient and scalable subspace clustering for very large data sets (mafia: Merging of Adaptive Finite Intervals)
- CBF (ACM Press02) – A new cell-based clustering method for large, high-dimensional data in data mining application
- CLTREE (ACM Press00) – Clustering through decision tree construction
- DOC (SIGMOD02) – A monte carlo algorithm for fast projective clustering

8/3/2006

PAKDD2006

74

Top-Down Subspace Clustering

- PROCLUS (SIGMOD 99)
- ORCLUS (SIGMOD00)
- FINDIT (Woo thesis02)
- δ -Clusters (DE Proc.02)
- COSA (J.RoyalSS04)

8/3/2006

PAKDD2006

75

PROCLUS

- PROCLUS (SIGMOD 99) – Fast algorithms for projected clustering
 - Modification of k-means algorithm
 - A top-down algorithm by finding *dense* regions into different subspaces
 - User inputs the number of clusters and average cluster dimensionality (unrealistic for real-world data sets)
 - Use cluster centers and points near to it to compute statistics. These determine the *relevant* cluster dimensions of the clusters

8/3/2006

PAKDD2006

76

PROCLUS (Cont'd)

- PROCLUS (SIGMOD 99) – Modification of k-means algorithm
 - Initialization
 - Greedy algorithm to select potential medoids that are far apart from each other
 - Iteration
 - Select a random set of medoids from the reduced dataset
 - Replace bad medoids to improve the clustering
 - Cluster refinement
 - Compute new dimensions for each medoid
 - Reassign points to medoids, removing outliers

8/3/2006

PAKDD2006

77

Others

- ORCLUS (SIGMOD00) – Finding generalized projected clusters in high dimensional spaces
- FINDIT (Woo thesis02) – a Fast and Intelligent Subspace Clustering Algorithm using Dimension Voting
- δ -Clusters (DE Proc.02) – Capturing subspace correlation in a large data set
- COSA (J.RoyalSS04) – Clustering objects on subsets of attributes

8/3/2006

PAKDD2006

78

Problems of Existing Subspace Clustering

- Computational complexity
 - CLIQUE splits each dimension into several CDUs (candidate dense unit) ---> large complexity
- Sensitive to the input parameters
 - CLIQUE needs to carefully set the parameters to determine the percentage of relevant data points
 - PROCLUS asks user to specify the average dimension of the cluster
- Hard determination of dimensions

8/3/2006

PAKDD2006

79

New Subspace Clustering for Text Documents

- Dimension weighting
- Soft determination of dimensions
- SKWIC --- Simultaneous keyword identification and clustering of text documents
- **ASI --- Adaptive Subspace Iteration**
- **FWKM --- Feature Weighting k-Means**

8/3/2006

PAKDD2006

80

SKWIC

- Simultaneous keyword identification and clustering of text documents
 - Clustering documents
 - Based on the k-means clustering algorithm
 - Cluster-dependent keyword weighting
 - Not all terms are considered equally relevant in a single category of text documents
 - Assign different weights to a term in different clusters
- Learn a different set of term weights for each cluster in an unsupervised manner

8/3/2006

PAKDD2006

81

SKWIC (Cont'd)

- Hard clustering

$$J(C, V; X) = \sum_{i=1}^C \sum_{x_j \in X_j} \sum_{k=1}^m v_{ik} D_{wc_{ij}}^k + \sum_{i=1}^C \delta_i \sum_{k=1}^m v_{ik}^2$$

- C is the number of cluster centers
- $V=[v_{ik}]$ is the set of relevance weights of keyword k in cluster i
 - where $v_{ik} \in [0,1], \forall i, k$
 - and $\sum_{k=1}^m v_{ik} = 1, \forall i$
- X is the document collection
- m is the number of dimensions

8/3/2006

PAKDD2006

82

SKWIC (Cont'd)

$$J(C, V; X) = \sum_{i=1}^C \sum_{x_j \in X_j} \sum_{k=1}^m v_{ik} D_{wc_{ij}}^k + \sum_{i=1}^C \delta_i \sum_{k=1}^m v_{ik}^2$$

$$D_{wc_{ij}}^k = \frac{1}{n} - (x_{jk} \cdot c_{ik})$$

- δ_i is very important in the algorithm since it reflects the importance of the second term relative to the first term
- δ_i should be chosen such that both terms are of the same order of magnitude

8/3/2006

PAKDD2006

83

SKWIC (Cont'd)

- Soft clustering

$$J(C, U, V; X) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^c \sum_{k=1}^m v_{ik} D_{wc_{ij}}^k + \sum_{i=1}^C \delta_i \sum_{k=1}^m v_{ik}^2$$

- $U=[u_{ij}]$ is the optimal soft partitioning membership
 - where $u_{ij} \in [0,1], \forall i, j$
 - and $0 < \sum_{j=1}^N u_{ij} < N, \forall i, j$
 - and $\sum_{i=1}^C u_{ij} = 1, \forall j$

8/3/2006

PAKDD2006

84

Problems of SKWIC

- Optimization problems
- Update unknowns in each step
- Formula for updating cluster centers
- SKWIC is dependent on a parameter
 - Depend on parameter δ_i
 - Not easy to find proper values for δ_i
 - Not guarantee the convergence of the optimization process
 - Update cluster centers/update the parameters?

8/3/2006

PAKDD2006

85

New Subspace Clustering for Text Documents

- SKWIC
- ASI
- FWKM

8/3/2006

PAKDD2006

86

ASI-Adaptive Subspace Iteration

- Tasks:
 - Data reduction
 - Assigning data points into clusters
 - Updating clusters based on the identified new subspace structures
 - Subspace identification
 - Identifying a subspace structure for each cluster from current cluster partitions

8/3/2006

PAKDD2006

87

ASI (Cont'd)

- Advantages
 - Mutually reinforcing the optimization procedure to exploit the duality of the data and terms
 - Deciding the number of clusters
 - Providing interpretable description (**meaningful word sets**) corresponding to each document cluster to represent their contents
- Present situation
 - Experiments only on binary VSM

8/3/2006

PAKDD2006

88

New Subspace Clustering for Text Documents

- SKWIC
- ASI
- FWKM

8/3/2006

PAKDD2006

89

FWKM (Feature Weighting k-Means)

- Recall
 - Characteristics of text data
 - Large volume
 - High dimensionality
 - Sparsity
 - Requirements on clustering techniques
 - Scalable to very large and high dimensional data
 - Able to find clusters from subspaces, i.e., the subsets of keywords in the text dataset

8/3/2006

PAKDD2006

90

FWKM (Cont'd)

- Motivation
 - K-means algorithm and subspace clustering
 - scalable
 - subspace
 - Understandable representation of clusters
 - Each subset of keywords for each cluster

8/3/2006

PAKDD2006

91

FWKM (Cont'd)

- Subspace clustering seeks clusters from subspaces of very high dimensional data [Parsons et al., 2004]
- Feature weighting k-means clustering [Chan et al., 2004]

$$F(W, Z, \Lambda) = \sum_{i=1}^k \sum_{j=1}^n \sum_{l=1}^m w_{i,j} \lambda_{i,j}^\beta d(z_{i,j}, x_{j,l})$$

- Subspace clustering of text documents with feature weighting k-means algorithm [Jing et al., 2005 PAKDD]

$$F(W, Z, \Lambda) = \sum_{i=1}^k \sum_{j=1}^n \sum_{l=1}^m w_{i,j} \lambda_{i,j}^\beta [d(z_{i,j}, x_{j,l}) + \sigma]$$

- The parameter sigma is to tackle the case of degeneration/baseline

8/3/2006

PAKDD2006

92

FWKM --- Algorithm

- Step 1: Initialize the cluster centers Z and weight matrix Λ
- Step 2: Use distance functions to get the partition matrix W
- Step 3: Update the cluster centers Z
- Step 4: Calculate the feature weights Λ

Repeat the Step 2, 3, and 4 until there is no improvement for the objective function

8/3/2006

PAKDD2006

93

FWKM --- Feature Weighting

- Weight Calculation

$$\lambda_{i,j} = \frac{1}{\sum_{j=1}^m \left[\sum_{l=1}^n \tilde{w}_{i,j} [d(\tilde{z}_{i,j}, x_{j,l}) + \sigma] \right]^{1/(\beta-1)}} \quad (1)$$

- E.g.

	t_0	t_1	t_2	t_3	t_4	
C_0	x_0	1	2	3	0	2
	x_1	2	3	1	0	2
C_1	x_2	0	0	1	3	2
	x_3	0	0	2	1	3

8/3/2006

PAKDD2006

94

FWKM --- Calculation σ

- Formula

$$\sigma = \frac{\sum_{j=1}^n \sum_{l=1}^m d(x_{j,l}, o_i)}{\tilde{n} \cdot m} \quad (2)$$

- use a sample instead of the entire data
- Reason
 - the value of σ affects the feature weighting
 - much larger --- ignore the feature values
 - too smaller --- turn back to standard k-means

8/3/2006

PAKDD2006

95

Experiments on Real Datasets

- Text data 20-Newsgroups - DSI

DataSet	Source	n_d	DataSet	Source	n_d
A2	alt.atheism	100	B2	talk.politics.mideast	100
	comp.graphics	100		talk.politics.misc	100
A4	comp.graphics	100	B4	comp.graphics	100
	rec.sport.baseball	100		comp.os.ms-windows	100
	sci.space	100		rec.autos	100
	talk.politics.mideast	100		sci.electronics	100
A4-U	comp.graphics	120	B4-U	comp.graphics	120
	rec.sport.baseball	100		comp.os.ms-windows	100
	sci.space	59		rec.autos	59
	talk.politics.mideast	20		sci.electronics	20

8/3/2006

PAKDD2006

96

Experiments on Real Datasets

Text data 20-Newsgroups – DS2

Dataset	n	m	k	Dataset	n	m	k	Dataset	n	m	k
D_1	15905	500	20	E_1	2000	1100	20	F_1	1500	500	3
D_2	15905	800	20	E_2	4000	1100	20	F_2	1500	500	5
D_3	15905	1100	20	E_3	8000	1100	20	F_3	1500	500	7
D_4	15905	1300	20	E_4	15905	1100	20	F_4	1500	500	10
D_5	15905	1700	20					F_5	1500	500	12
D_6	15905	2000	20								

8/3/2006

PAKDD2006

97

Experimental Results on DS1

Comparison on clustering quality

	<i>Bi-K-Means</i>	<i>FWKM</i>	<i>PROCLUS</i>	<i>HARP</i>	<i>COS4</i>	<i>SCADJ</i>
A_2	0.2146	0.2057	0.5254	0.5010	0.9999	0.2777
	0.9650	0.9599	0.7190	0.8884	0.5781	0.9490
	0.7857	0.7961	0.2334	0.4984	0.0008	0.7226
B_2	0.5394	0.4014	0.8395	0.5562	0.9973	0.5664
	0.8800	0.9043	0.6604	0.8020	0.5413	0.8661
	0.4706	0.6050	0.0789	0.0299	0.0027	0.4260
A_4	0.1919	0.2509	0.5548	0.7671	0.9902	0.4214
	0.9376	0.9003	0.6450	0.5073	0.3152	0.8383
	0.8083	0.7554	0.2909	0.2023	0.0099	0.5854
B_4	0.6193	0.3574	0.7291	0.8933	0.9819	0.5380
	0.7049	0.8631	0.4011	0.3840	0.3621	0.7711
	0.3822	0.6467	0.0791	0.0538	0.0236	0.4174
A_4-U	0.2830	0.1513	0.7342	0.8389	0.8768	0.5286
	0.8961	0.9591	0.5239	0.4819	0.4159	0.8719
	0.7126	0.8480	0.1867	0.1688	0.0187	0.5563
B_4-U	0.5357	0.2314	0.5758	0.9535	0.8614	0.3591
	0.6586	0.9205	0.5739	0.3364	0.3599	0.8597
	0.3793	0.7385	0.1684	0.0250	0.0300	0.6442

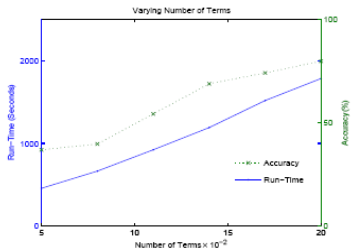
8/3/2006

PAKDD2006

98

Experimental Results on DS2 -1

Scalable to the number of features/terms



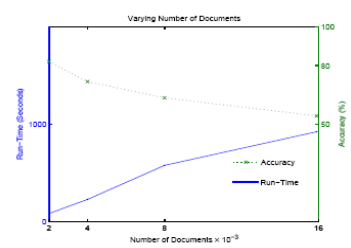
8/3/2006

PAKDD2006

99

Experimental Results on DS2 -2

Scalable to the number of objects



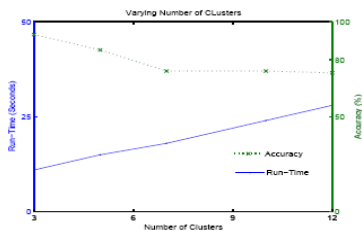
8/3/2006

PAKDD2006

100

Experimental Results on DS2 -3

Scalable to the number of clusters



8/3/2006

PAKDD2006

101

Text Clustering Methods

- Subspace clustering
- Semi-supervised Clustering

8/3/2006

PAKDD2006

102

Why Use Semi-supervised Clustering?

- Motivation
 - Large amounts of unlabeled data
 - More is being produced all the time
 - Expensive to generate labels for data
 - Usually requires human intervention
 - Want to use limited amounts of labeled data to guide the clustering of the entire dataset in order to provide a better clustering result

8/3/2006

PAKDD2006

103

What is Semi-supervised Clustering?

- Use human input to provide knowledge for part of the data
 - Improve existing naive clustering methods
 - Use labeled data to guide clustering of unlabeled data
 - End result is a better clustering of data

8/3/2006

PAKDD2006

104

Methods of Semi-supervised

- Seed-Based:
 - Initializing seeds with labeled data
 - Constraint-Based:
 - Pairwise constraints of Must-link, or Cannot-link labels
 - Set M of must link constraints
 - Set C of cannot link constraints
 - A list of associated costs for violating Must-link or cannot-link requirements (modify the objective function)
- $$J = \sum_{x_i \in X} \|x_i - \mu_{l_i}\|^2 + \sum_{(x_i, x_j) \in M} w_{ij} I[l_i \neq l_j] + \sum_{(x_i, x_j) \in C} \bar{w}_{ij} I[l_i = l_j]$$
- Class labels do not have to be known, but a user can still specify relationship between points

8/3/2006

PAKDD2006

105

Methods of Semi-supervised (Cont'd)

- Feedback-based
 - Include user feedback to direct the clustering process
 - Interactively evaluate the quality of clustering results
- Topic-based
 - Topic is not in format of labeled data
 - Combine documents with topic by optimization unsupervised process and supervised process
 - Weighted scheme
 - Normalized scheme
 - Hybrid scheme

8/3/2006

PAKDD2006

106

Important References

- S. Basu, M. Bilenko, and R. Monney, A probabilistic framework for semi-supervised clustering, Proc. Of 10th international conference on knowledge discovery and data mining, 2004
- A. Becks and C. Seeling, SWAPIT: a multiple views paradigm for exploring associations of texts and structured data, Proc. Of the working conf. on advanced visual interfaces, 193-196, 2004
- H. Frigui and O. Nasraoui, Simultaneous clustering and dynamic keyword weighting for text documents, Survey of text mining, Michael Berry, 45-70, 2004
- J. Han, and M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann, 2001
- A.K. Jain, M.N. Murty, and P.J. Flynn, Data clustering: A review, ACM Computing Surveys, 31(3), 264-323, 1999
- L. Jing, M.K. Ng, J.Z. Huang and J. Xu, Subspace clustering of text documents with feature weighting k-means algorithm, PAKDD, 802-812, 2005

8/3/2006

PAKDD2006

107

Important References (Cont'd)

- H. Kim and S. Lee, A semi-supervised document clustering technique for information organization, In Proc. of the 9th international conference on information and knowledge management, 30-37, 2000
- T. Li, S. Ma, and M. Ogihara, Document clustering via adaptive subspace iteration, SIGIR, 218-225, 2004
- J. Liu, W. Wang, and J. Yang, A framework for ontology-driven subspace clustering, SIGKDD, 623-628, 2004
- L. Parsons, E. Haque, and H. Liu, Subspace clustering for high dimensional data: a review, SIGKDD, 6(1) 90-105, 2004
- Y. Zhao and G. Karypic, Topic-driven clustering for document datasets, SIAM-05

8/3/2006

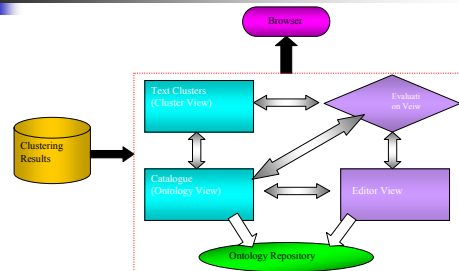
PAKDD2006

108

Section 5

Post-processing Techniques with Ontology and Clustering Visualization

Block Diagram



8/3/2006

PAKDD2006

110

Visualization Methods

- Cluster View
- Ontology View
- Evaluation View
- Editor View

8/3/2006

PAKDD2006

111

Cluster View

- Cluster View
 - The module aims at helping the user to explore a set of text documents based on the document content.
 - It displays inter-document and inter-cluster associations based on a measure of similarity between each pair of documents or pair of clusters, therefore visualizing the cluster structure of the document space.
- Example
 - RefViz

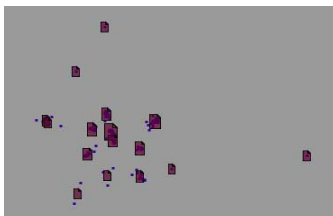
8/3/2006

PAKDD2006

112

Example

- A document map approach (RefViz)



8/3/2006

PAKDD2006

113

Ontology View

- Ontology View
 - The module enables the user to navigate through document collections by means of domain-specific topic catalogues.
 - Domain catalogues can be defined by different types of ontologies, e.g., taxonomies or topic maps.
- Example
 - Visivimo

8/3/2006

PAKDD2006

114

Section 6

Text Clustering Applications and Systems

Applications of Text Clustering

- Query Routing
- Cluster-based Browsing
- Result Set Clustering
- Result Set Expansion
- Query Refinement
- Bio-informatics Application

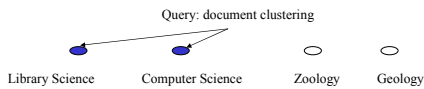
8/3/2006

PAKDD2006

122

Query Routing

- Documents distributed in several information servers
 - **Relevant documents** are clustered and put in one or proximate servers
 - Generating description to represent all of documents in a cluster
- When retrieving information
 - **Identifying relevant clusters** based on the relevance between queries and description of clusters
 - Forwarding queries to the corresponding servers
 - Merging the results
- An example



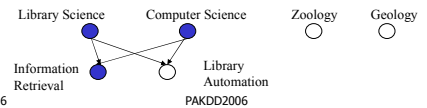
8/3/2006

PAKDD2006

123

Cluster-based Browsing

- The problems of expressing a vague information need as a formal query
- Scatter/Gather (Cutting, et. al., SIGIR'92)
 - Clustering documents into **topic-coherent groups**
 - Presenting **descriptive summaries** of the clusters to users
 - Users can browse and determine possible clusters hierarchy
 - Documents in the selected clusters are clustered and summaries are generated
 - Finally, documents are retrieved
- Vivisimo (www.vivisimo.com)
- An example



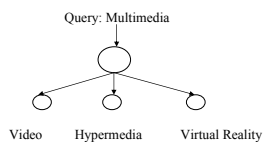
8/3/2006

PAKDD2006

124

Result Set Clustering

- Users' queries are often very short (about 1-3 words)
 - The result set contains relevant documents and irrelevant documents
- Clustering documents in the result set according to the degree of relevance
 - Helping users figure out their real information needs
 - Easily retrieving relevant documents
- An example



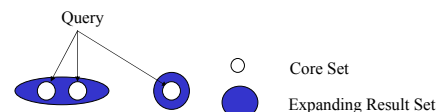
8/3/2006

PAKDD2006

125

Result Set Expansion

- Relevant documents may not match the input queries well
- Clustering relevant documents based on sophisticated features and clustering algorithms in data preparing phase
- Retrieving a **core set of documents** that match the query
- **Expanding** the results with documents not matching the query but **clustered with the documents in the core set**



8/3/2006

PAKDD2006

126

Query Refinement

- Terms in queries may not match the information needs of users
- Dynamically computing and suggesting **recall- and precision-enhancing terms** for a given query
- Term suggestion
 - Grouping retrieved documents into topic-cohesive clusters
 - Terms in the core documents: general concepts
 - Terms in the margin documents: specific concepts

8/3/2006

PAKDD2006

127

Bio-informatics Application

- Extract relations between biological entities (e.g., protein-protein interactions)
- Update biomedical databases (largely manual)
- Explain the underlying biological mechanisms associated with a cluster of genes
- Drug discovery

8/3/2006

PAKDD2006

128

Existing Text Clustering Systems

- UIMA - IBM
- KAON - AIFB
- GATE – NLP group in Univ. of Sheffield
- BOW – McCallum in CMU

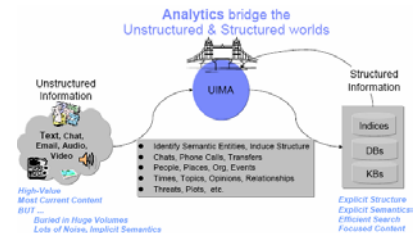
8/3/2006

PAKDD2006

129

UIMA - IBM

- IBM's UIMA (**unstructured information management architecture**) is an architecture and software framework that helps build the bridge between unstructured and structured worlds.



8/3/2006

130

UIMA Techniques

- IBM's UIMA supports creating, discovering, composing and deploying a broad range of analysis capabilities and linking them to structured information services.
- UIMA applications make use of a variety of analysis technologies including:
 - Statistical and rule-based Natural Language Processing (NLP)
 - Information Retrieval (IR)
 - Machine Learning
 - Ontologies
 - Automated reasoning
 - Knowledge Source (e.g., WordNet)

8/3/2006

PAKDD2006

131

KAON - AIFB

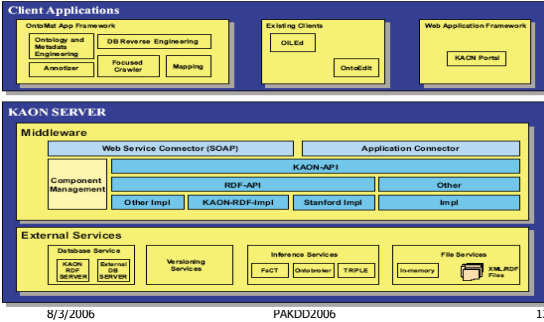
- KAON is an open-source ontology management infrastructure for semantics-driven business applications
- KAON includes a comprehensive tool suite allowing easy ontology management and application.
- KAON architecture has three layers (shown as the figure in next slide):
 - Client layer
 - Middleware layer
 - External services layers

8/3/2006

PAKDD2006

132

KAON – AIFB



GATE

- GATE stands for **g**eneral **a**rchitecture for **t**ext **e**ngineering.
- GATE combines human language computation with software engineering.
- GATE constitutes an infrastructural system that supports research and development of language processing software.

8/3/2006

PAKDD2006

134

GATE

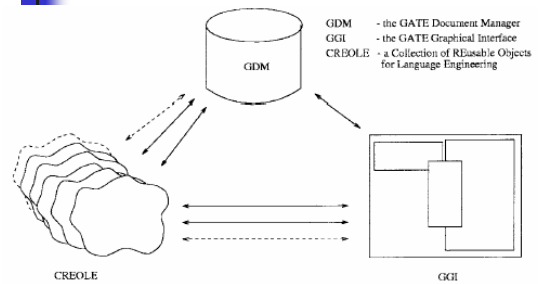
- GATE includes three principal elements:
 - GDM (**g**ate **d**ocument **m**anager), based on the existing document manager (TIPSTER)
 - CREOLE: a collection of **r**eusable **o**bjects for **l**anguage **e**ngineering, a set of LE components integrated with the system
 - GGI: the **g**ate **g**raphical interface, a development tool for LE R&D, providing integrated access to the services of the other components and adding visualization and debugging tools

8/3/2006

PAKDD2006

135

GATE Architecture



BOW

- BOW is a library of *C* code useful for writing statistical text analysis, language modeling and information retrieval programs
- The current distribution includes the library:
 - document classification (rainbow)
 - document retrieval (arrow)
 - document clustering (crossbow)

8/3/2006

PAKDD2006

137

Important References

- BOW, <http://www.cs.cmu.edu/mccallum/bow>, CMU, 1996
- GATE, gate.ac.uk/gate, Univ. of Sheffield, UK, 2002
- KAON, kaon.semanticweb.org, AIFB, German, 2002
- SPSS, Meeting the challenge of Text, White paper, SPSS Inc. LQWP-1203, 2003
- SyNTHEMA, www.synthema.it/tewat, Lexical systems Lab., Italy
- UIMA, www.research.ibm.com/UIMA, IBM, 2004
- Vivisimo, www.vivisimo.com
- H. Cunningham, A general architecture for text engineering, *Computers and the humanities*, 36, 223-254, 2002

8/3/2006

PAKDD2006

138



Important References

- D. Ferrucci, and A. Lally, An architectural approach to unstructured information processing in the corporate research environment, *Natural language engineering* 10(3/4), 327-348, 2004
- P. Kankar, S. Adak, A. Sarkar, K. Murari, and G. Sharma, MedMeSH summarizer: text mining for gene clusters, In *Proc. of the second SIAM international conference on data mining*, Arlington, VA, USA, 2002
- SAS, *Getting started with SAS 9.1 Text Miner*, SAS Institute Inc., 2004
- D.R. Swanson, Medical literature as a potential source of new knowledge, *Bull Med Libr Assoc*, 78(1), 1990

8/3/2006

PAKDD2006

139



Summary

- Text clustering is an important technology to improve efficiency and effectiveness in information retrieval
 - Possible applications are wide
- Techniques of text clustering
 - Extraction of features to represent documents
 - Relevance functions between documents
 - Clustering algorithms
 - Visualization methods

8/3/2006

PAKDD2006

140



DataSets

- 20Newsgroups:
<http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html>
- Reuters-21578:
<http://www.daviddlewis.com/resources/testcollections/reuters21578/>
- WebKB: <http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.tar.gz>
- CSTR:
<http://www.cs.rochester.edu/trs>
- MEDLINE: http://www.nlm.nih.gov/databases/databases_medline.html

8/3/2006

PAKDD2006

141