



MLIA Architectures: Advanced Services Beyond Searching

Giorgio Maria Di Nunzio



IMS Research Group
Dept. of Information Engineering
University of Padua



- Part I
 - Cross-lingual Text Categorization

- Part II
 - Log Analysis

- Automated Text Categorization (ATC):
 - introduction and definition of the problem;
 - representation of documents for ATC;
- Probabilistic models for ATC:
 - Bernoulli Naïve Bayes model;
 - multinomial Naïve Bayes model
 - two-dimensional model
- Cross-lingual Text Categorization



What is Categorization?

- Categorization is the process of dividing the world into groups of entities whose members are in some way similar to each other.
- These groups of entities, or categories, are populated with members that share some perceptible similarity within a given context, that is the context may vary.





Ontology and Categories

- Ontology is the study of the basic categories of being and their relations (and more).
 - deals with questions concerning what entities exist or can be said to exist
 - how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences
- **Categories** (IV B.C.) is a text from Aristotle's “**Organon**” that enumerates all the possible kinds of thing which can be the subject or the predicate of a proposition.



Tree of Porphyry (IV A.D.)

Supreme genus:

Differentiae:

Subordinate genera:

Differentiae:

Subordinate genera:

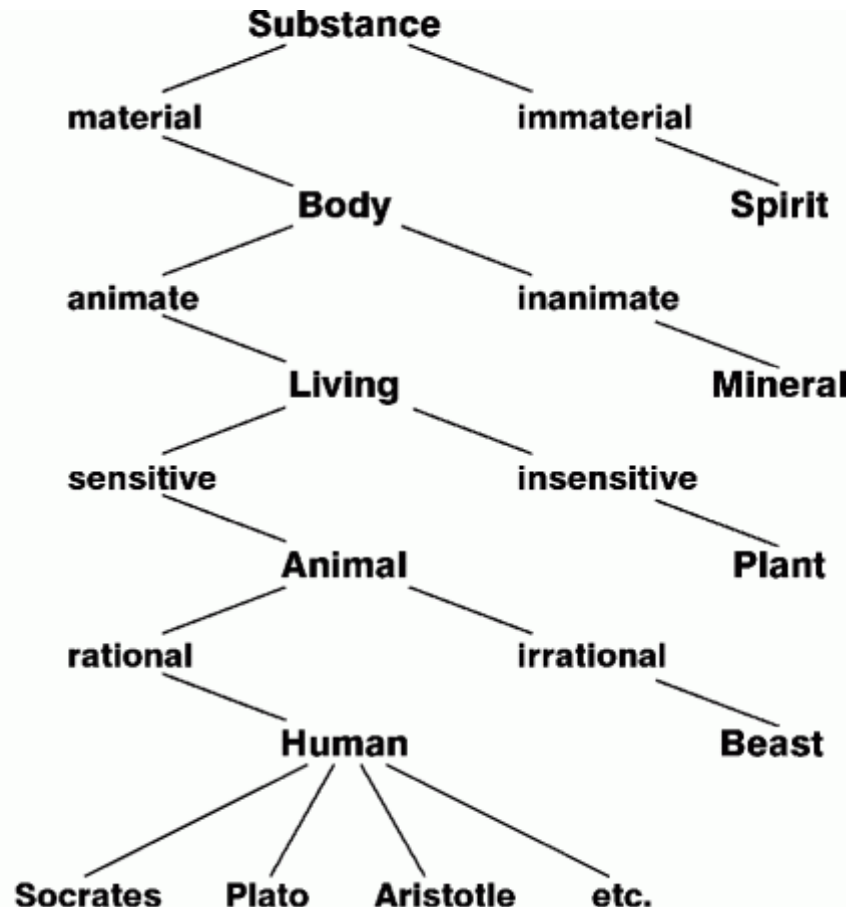
Differentiae:

Proximate genera:

Differentiae:

Species:

Individuals:



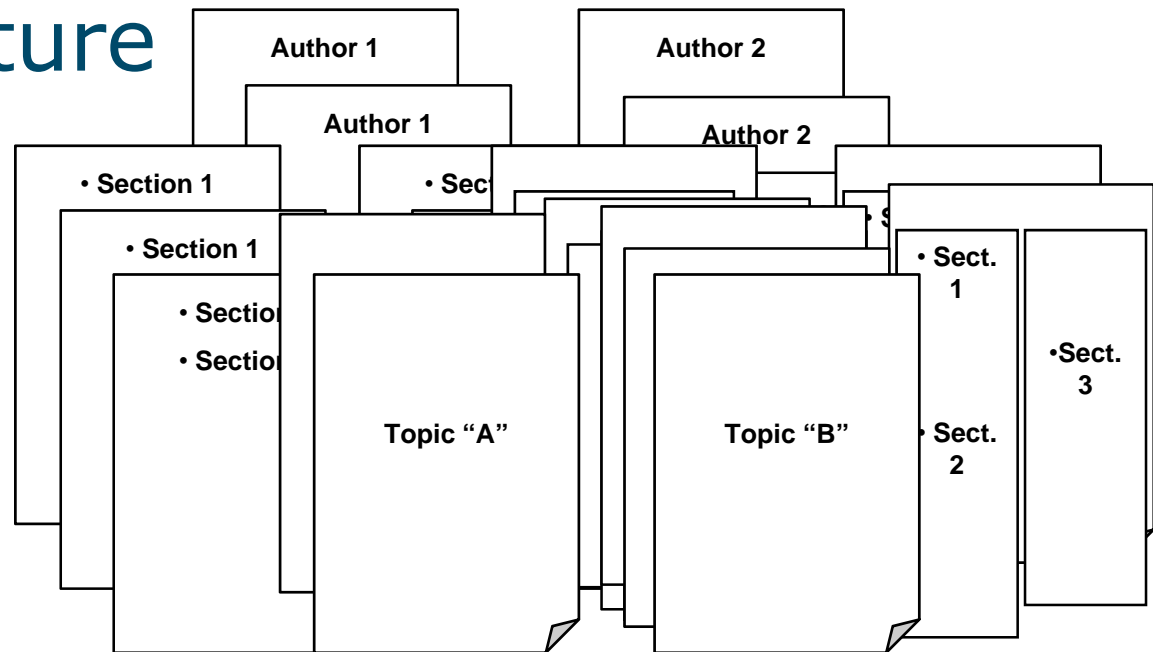


What is Text Categorization?

- When the entities to be categorized are textual documents, the process of organizing sets of documents into categories according to the property of the documents and to the essential properties of the category, is called *Text Categorization*.
- When dealing with textual documents, different views on these documents should be distinguished, because each different view underlines different properties for a document.



- External attributes
- Logical structure
- Layout structure
- Content





What is Automated Text Categorization (ATC)?

- A general inductive process, the learner, automatically builds a categorizer for the categories involved in the process by observing the properties of a set of pre-classified documents (supervised learning).
- From these properties, the inductive process learns the characteristics that a new unseen document should have in order to be categorized under a category.



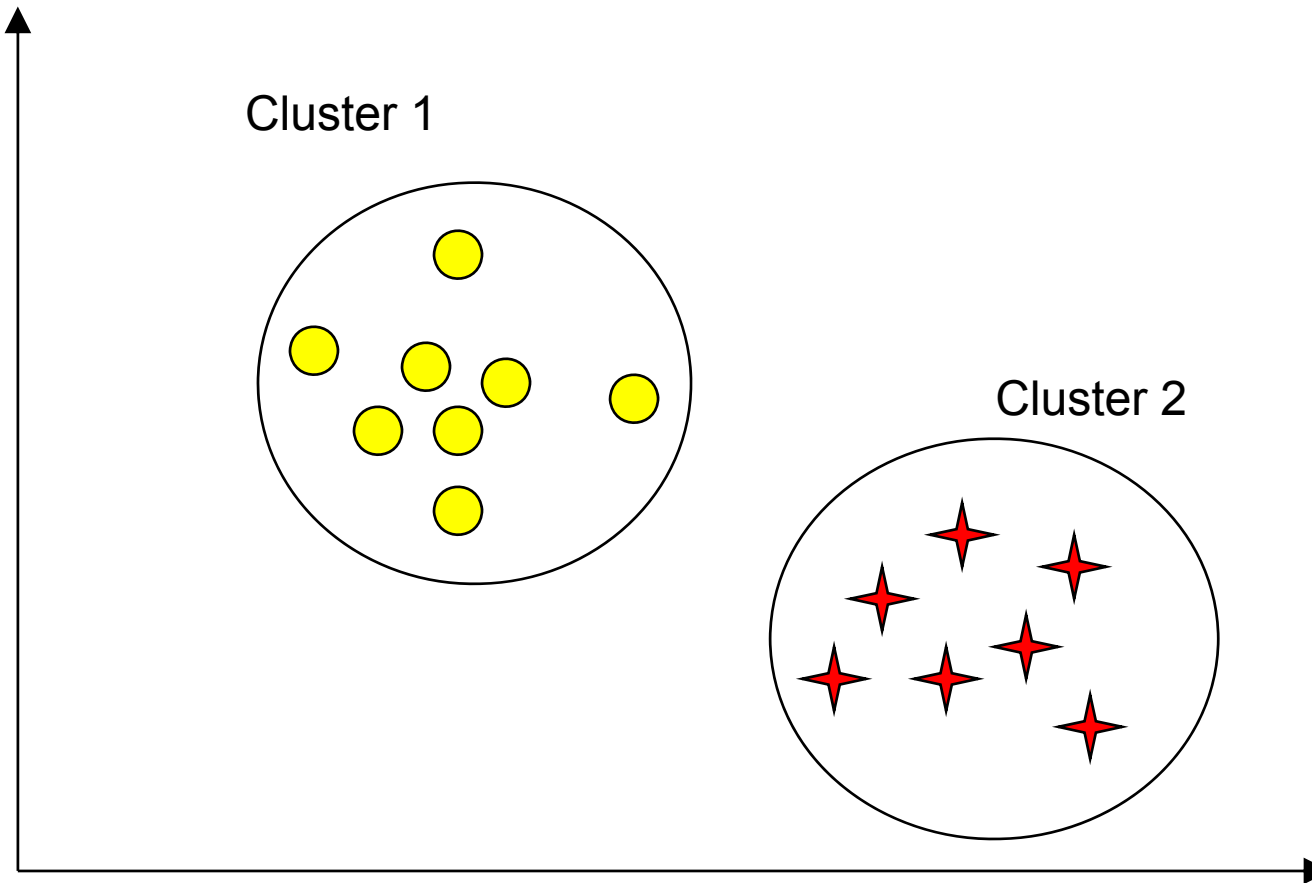


Machine Learning

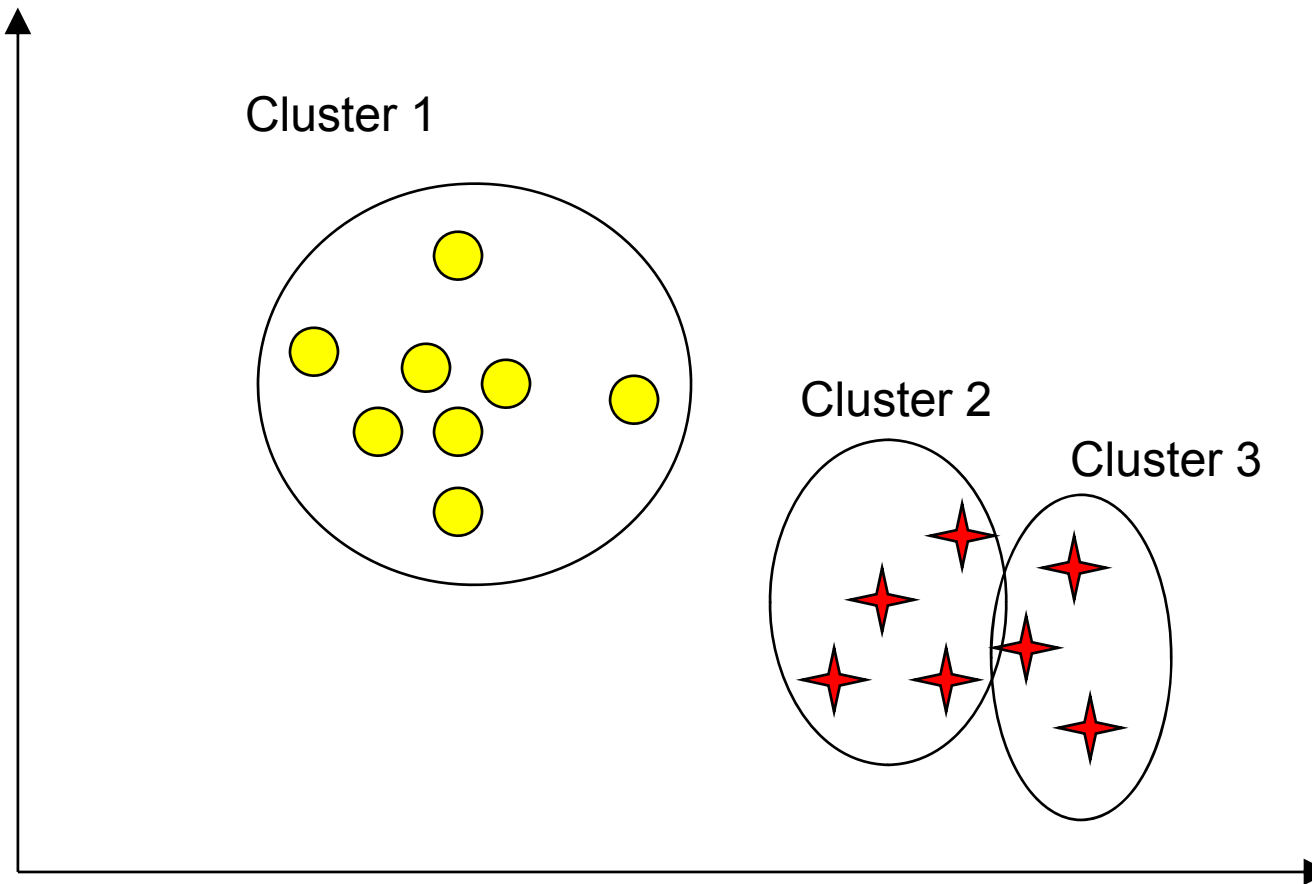
- Supervised Learning: given examples of inputs and corresponding desired outputs, predict on future inputs
- Unsupervised Learning: give only inputs, automatically discover representations, features, etc.



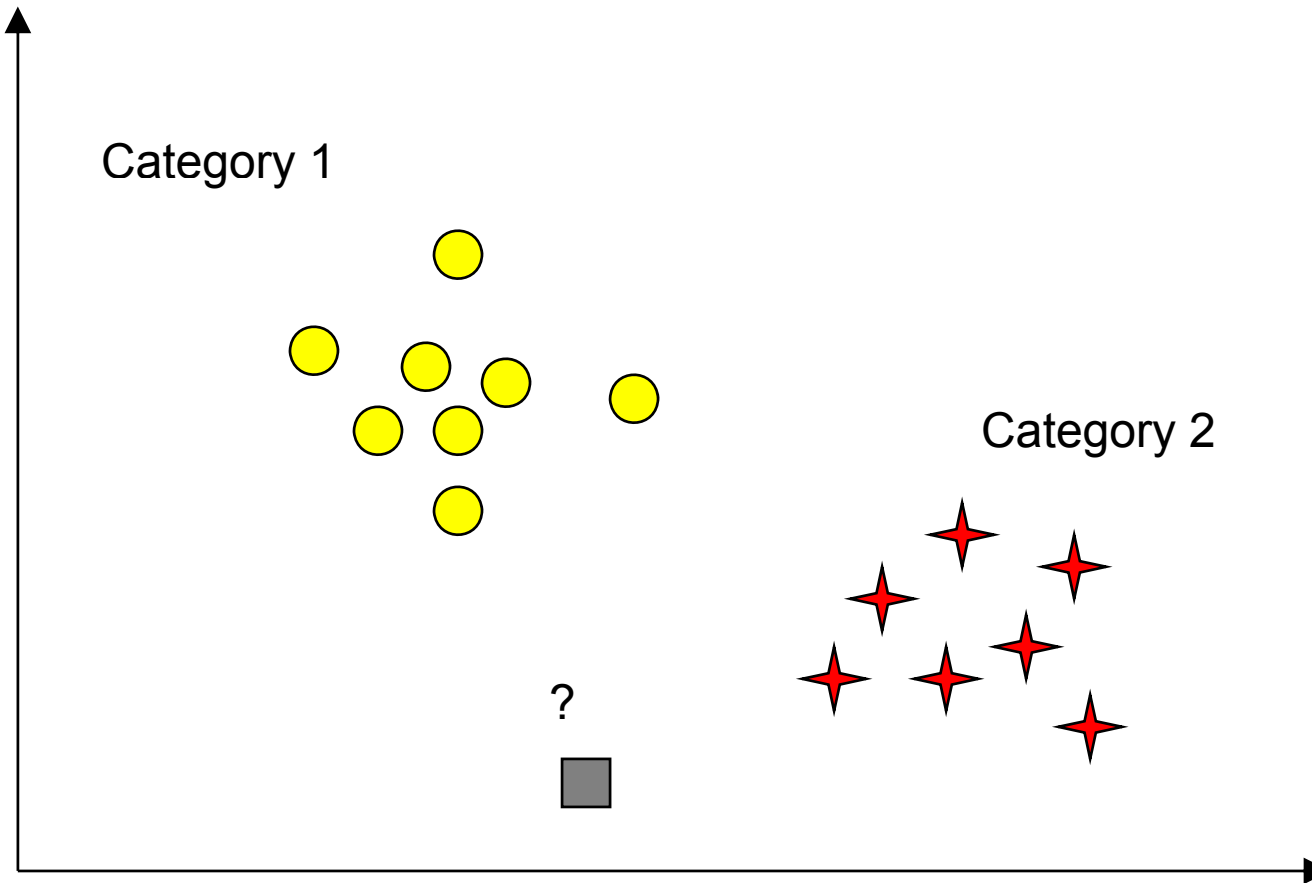
Categorization vs Clustering



Categorization vs Clustering



Categorization vs Clustering





Structure of Learning Machines

- Use a mathematical function $z = f(\mathbf{x})$
 - \mathbf{x} is the representation of the input
 - z is the representation of the output
- Hypothesis space and Parameters
 - $H : \{f(z|x, \theta)\}$
- How do we represent inputs/outputs and select the hypothesis space?





Representation of documents for ATC

- Set of terms that occur in the documents (*Bag-of-words approach*).
- A **vocabulary V** of term is needed to summarize and index the document content.
- The selection of V is made by successive refinements:
 - starting from an ideal/specific dictionary,
 - removing *stop-words* (articles, prepositions, conjunctions, etc.),
 - performing *stemming* (reduce terms to their roots),
 - selecting most valuable terms (*feature selection*).
- A document d is represented by numbers that represent the weight $w(\cdot)$ of each term that appears in the document.

$$d = \{ w(t_1), \dots, w(t_{|V|}) \}$$

- When weights are *non-binary*, they are usually computed by either statistical or probabilistic techniques.





Training vs Testing

- Training data: the X, Y we are given.
- Testing data: the X, Y we will see in future.
- Training error: the average value of loss on the training data.
- Test error: the average value of loss on the test data.
- What is our real goal? To do well on the data we have seen already?
- How to do this if we don't have test data?
- Are training and test documents sampled from the same distribution?





Probabilistic models for ATC

- Naïve Bayes model
 - Bernoulli model;
 - Multinomial model

- Visualization of Probabilistic Models
 - two-dimensional model

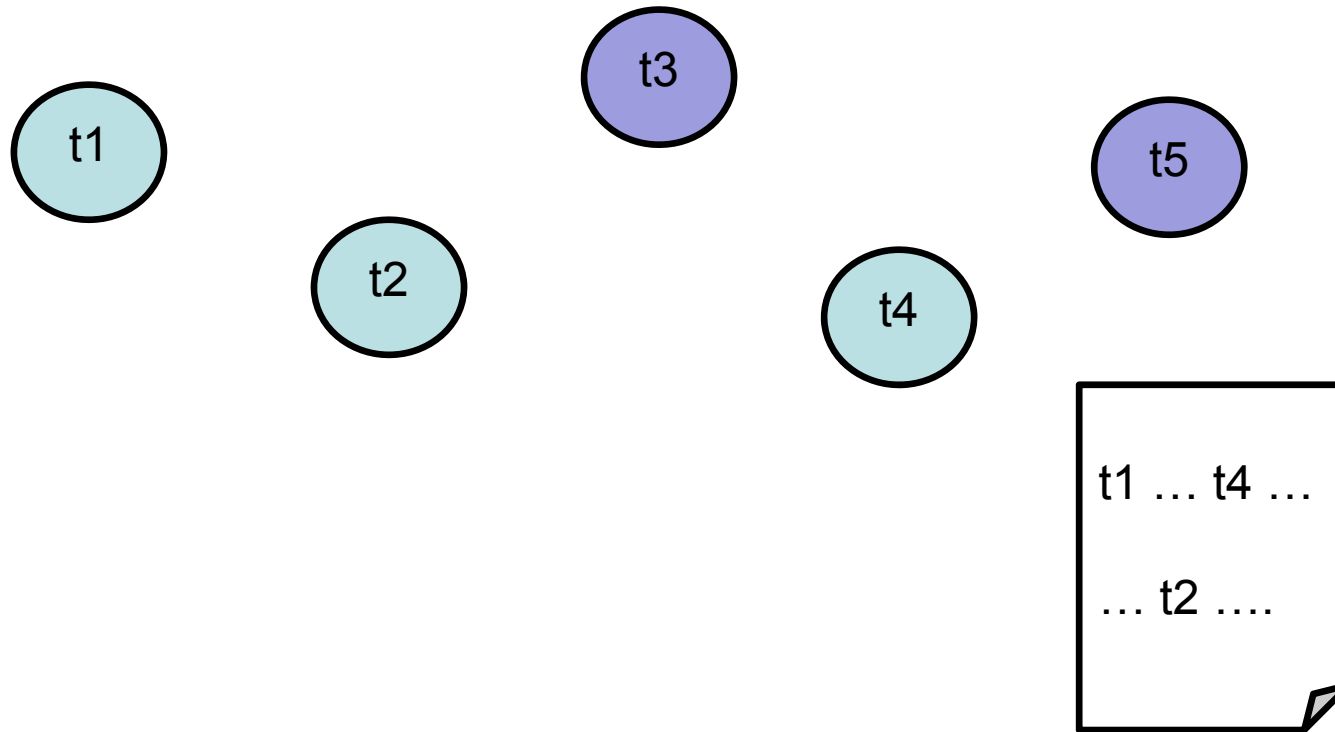


- Bayesian classifiers are probabilistic approaches that make strong assumptions about how the data is generated, and build a probabilistic model that embodies these assumptions.
 - Use Bayes' rule to estimate the probability $P(c|d)$
 - Need a collection of labeled training examples to estimate the probabilities of the model.

- Why Naïve?
$$P(c | d) = \frac{P(d | c)P(c)}{P(d)} ; \quad P(d | c) = \prod_t P(t | c)$$

- How to classify documents?
$$P(c | d) > P(\bar{c} | d)$$

Tossing Coins



Bernoulli NB Model

- A document is represented by a vector of binary attributes indicating which terms occur and do not occur in the document.
 - The number of times a term occurs in a document is not captured.

$$d = \{t_1, \bar{t}_2, t_3, t_4\}$$

- When calculating the probability of a document, you multiply the probability of all the attribute values, including the probability of non-occurrence for those terms that do not occur in the document.

$$P(d | c) = \prod_{k=1}^{|V|} P(t_k | c)^{x_k} (1 - P(t_k | c))^{1-x_k}$$





Learning Bernoulli NB Model

- Maximum Likelihood Estimates (MLE)
 - Use frequencies in the training data set
- Problems with MLE
 - What if we have no examples in the training data set?
 - How do we deal with zero probabilities?



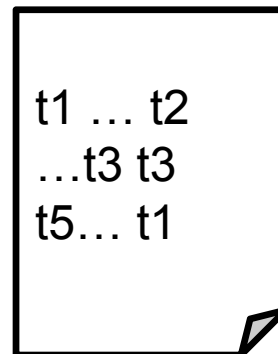
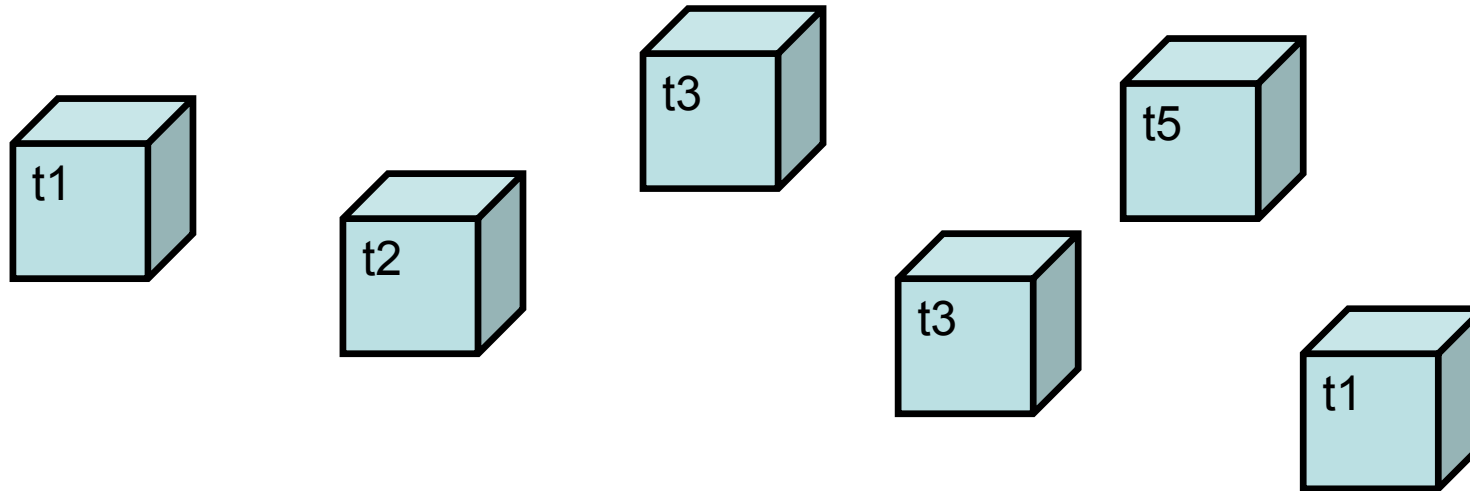


Smoothing probabilities

- Laplace smoothing
 - Add 1 to all frequency counts
 - Pros: easy to compute
 - Cons: give too much probability mass to unseen events
- Add “delta” to compute the extent of smoothing
- And many others...



Rolling Dice



t1 ... t2
...t3 t3
t5... t1

- A document is represented by a (multi-)set of terms that occur in the document.
 - The order of the terms is lost
 - the number of occurrences of each term in the document is captured.

$$d = \{t_3, t_1, t_3, t_3, t_4\}$$

- When calculating the probability of a document, only the probability of the terms that occur count.

$$P(d | c) = P(|d|) \frac{|d|!}{n_{t_1}! n_{t_2}! \dots n_{t_{|V|}}!} \prod_{k=1}^{|V|} P(t_k | c)^{n_{t_k}}$$





NB models: time complexity

- Training Time
- Test Time
- Linearly proportional
 - Just read in all the data



Any difference?

$$P(c | d) > P(\bar{c} | d)$$

$$\frac{P(c | d)}{P(\bar{c} | d)} > 1$$

$$\frac{P(d | c)P(c)}{P(d | \bar{c})P(\bar{c})} > 1$$

$$\frac{\prod_{k=1}^{|\mathcal{V}|} P(t_k | c)^{x_k} (1 - P(t_k | c))^{1-x_k} P(c)}{\prod_{k=1}^{|\mathcal{V}|} P(t_k | c)^{x_k} (1 - P(t_k | c))^{1-x_k} P(\bar{c})} > 1$$

$$\frac{\prod_{k=1}^{|\mathcal{V}|} P(t_k | c)^{n_{t_k}} P(c)}{\prod_{k=1}^{|\mathcal{V}|} P(t_k | \bar{c})^{n_{t_k}} P(\bar{c})} > 1$$

$$\frac{\prod_{k=1}^{|\mathcal{V}|} \left(P(t_k | c)^{x_k} (1 - P(t_k | c))^{1-x_k} \right) P(c)}{\prod_{k=1}^{|\mathcal{V}|} \left(P(t_k | \bar{c})^{x_k} (1 - P(t_k | \bar{c}))^{1-x_k} \right) P(\bar{c})} > 1$$

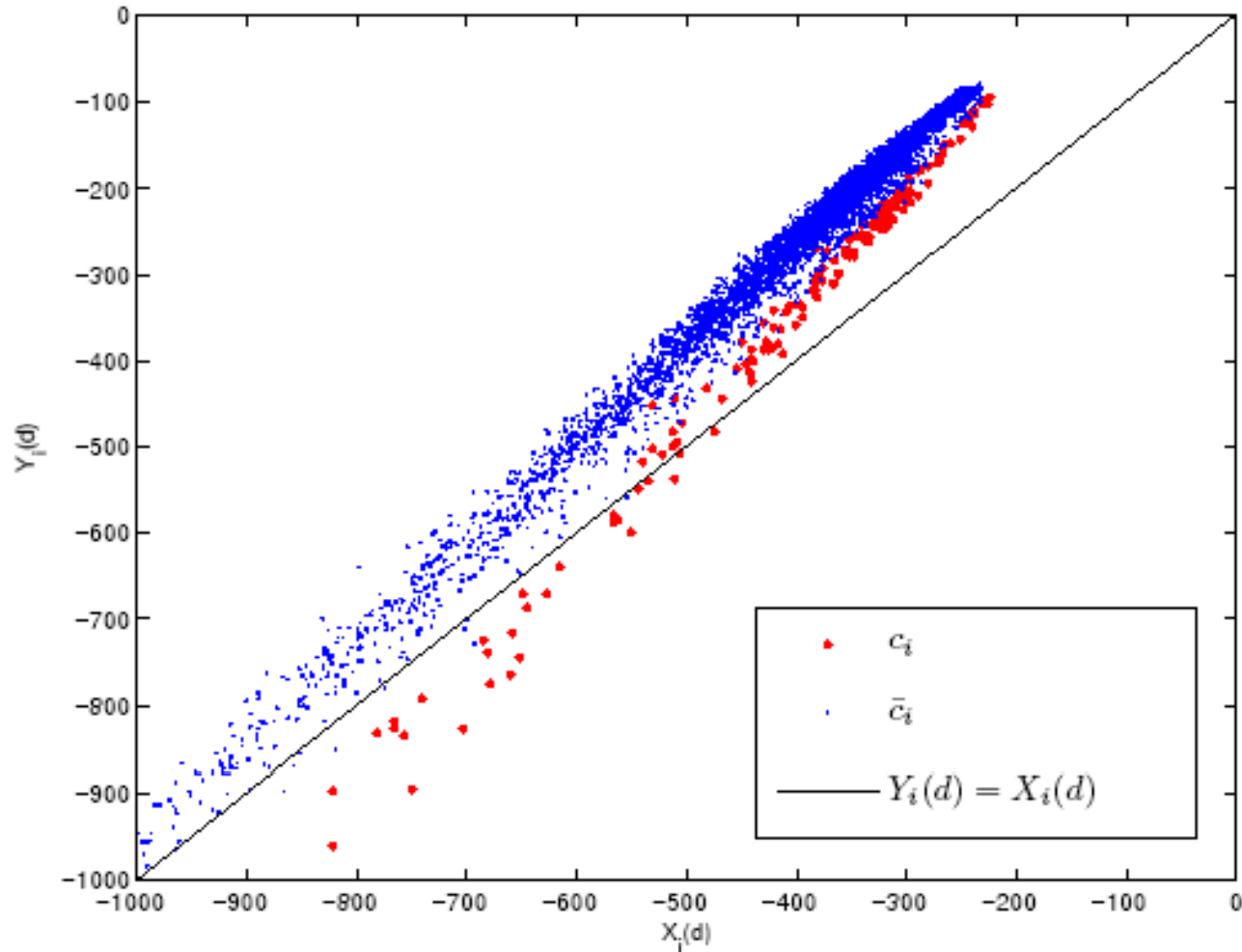
$$\frac{\prod_{k=1}^{|\mathcal{V}|} \left(\left(\frac{P(t_k | c)}{(1 - P(t_k | c))} \right)^{x_k} (1 - P(t_k | c)) \right) P(c)}{\prod_{k=1}^{|\mathcal{V}|} \left(\left(\frac{P(t_k | \bar{c})}{(1 - P(t_k | \bar{c}))} \right)^{x_k} (1 - P(t_k | \bar{c})) \right) P(\bar{c})} > 1$$

Definitely maybe!

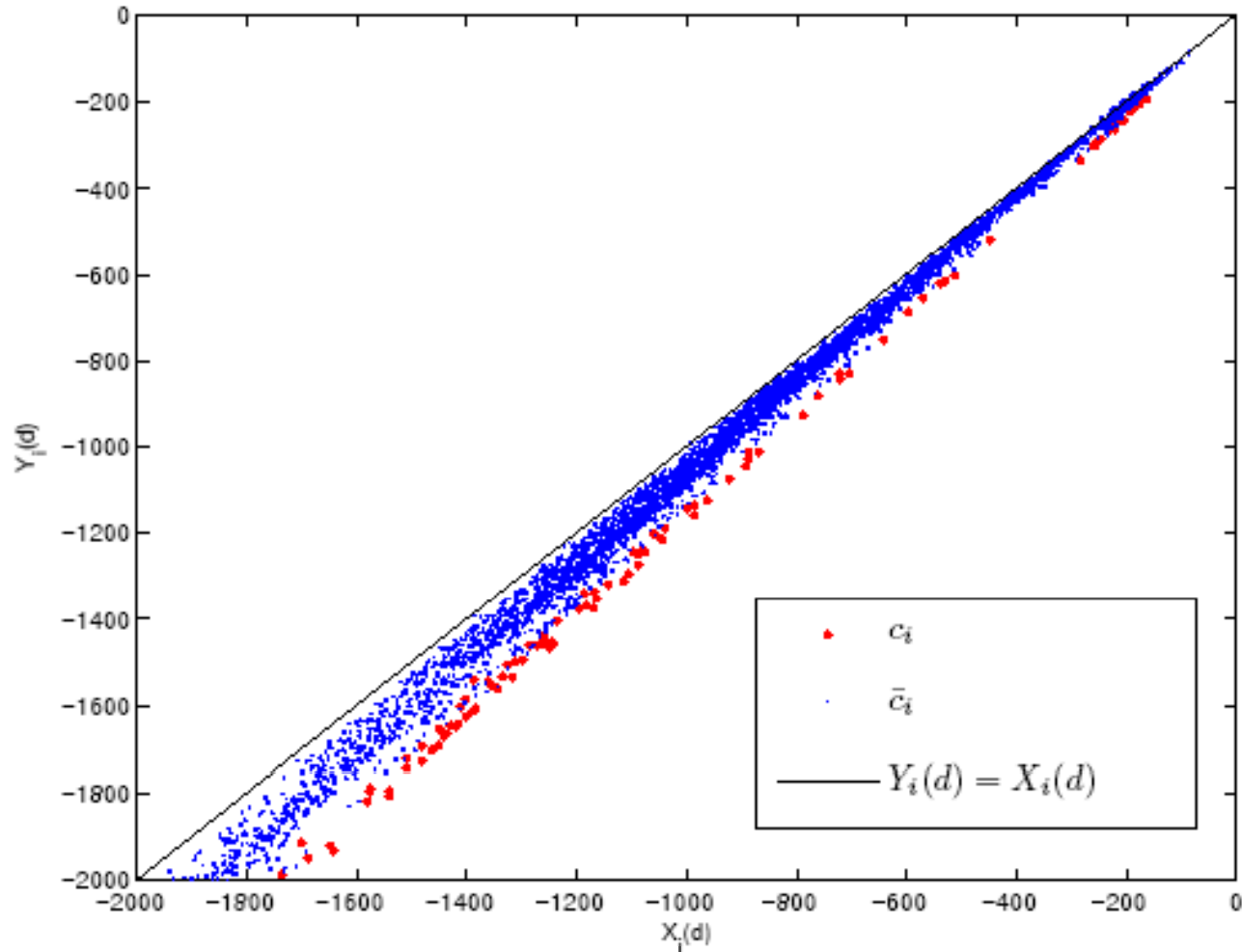
$$\prod_{k=1}^{|\mathcal{V}|} \left(\frac{P(t_k | c)(1 - P(t_k | \bar{c}))}{P(t_k | \bar{c})(1 - P(t_k | c))} \right)^{x_k} \prod_{k=1}^{|\mathcal{V}|} \frac{(1 - P(t_k | c)) P(c)}{(1 - P(t_k | \bar{c})) P(\bar{c})} > 1$$

- In the two-dimensional representation of documents, two coordinates $X_i(d)$ and $Y_i(d)$ are calculated for each document d of a category c_i .
- G. M. Di Nunzio, A. Micarelli, Pushing "Underfitting" to the Limit: Learning in Bidimensional Text Categorization, in: R. L. de Mántaras, L. Saitta (Eds.), Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), including Prestigious Applicants of Intelligent Systems (PAIS 2004), IOS Press, 2004, pp. 465-469.
- Rewrite log-odds
- G. M. Di Nunzio, Visualization and Classification of Documents: A New Probabilistic Model to Automated Text Classification, Bulletin of the IEEE Technical Committee on Digital Libraries (IEEE-TCDL) 2 (2).

2D model - Bernoulli

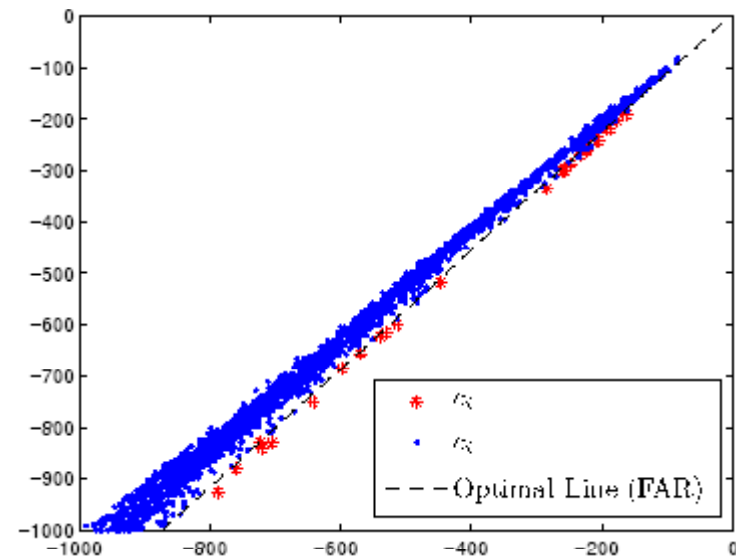
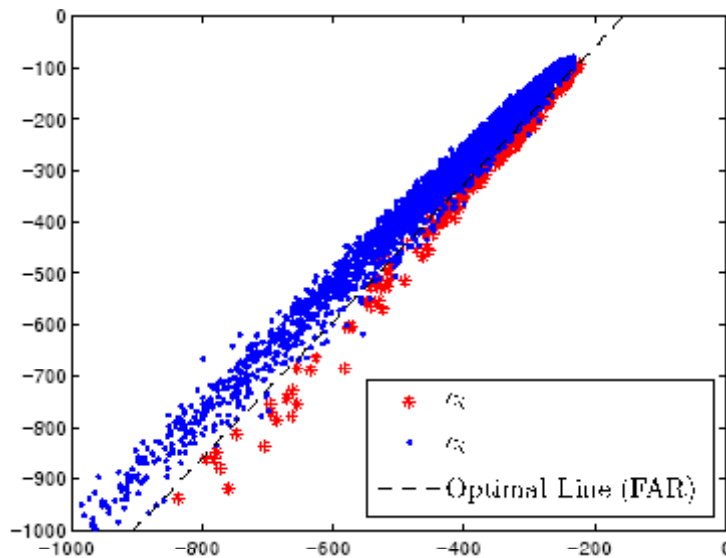


2D model - Multinomial



2D – Improving performance

- A better separation of the documents is reached when the plane is split into two regions
- G. Di Nunzio. Using Scatterplots to Understand and Improve Probabilistic Models for Text Categorization and Retrieval. International Journal of Approximate Reasoning, In Press 2009. <http://dx.doi.org/10.1016/j.ijar.2009.01.002>.



- Motivations
 - Extend ATC system in order to classify documents in different languages
- Most applications of TC, e.g. digital libraries, news wire filtering as well as Web page and e-mail categorization, are interesting applications of multilingual text classification (MLTC), where documents given in different languages are to be classified by topic or similar criteria.

CLTC vs CLIR

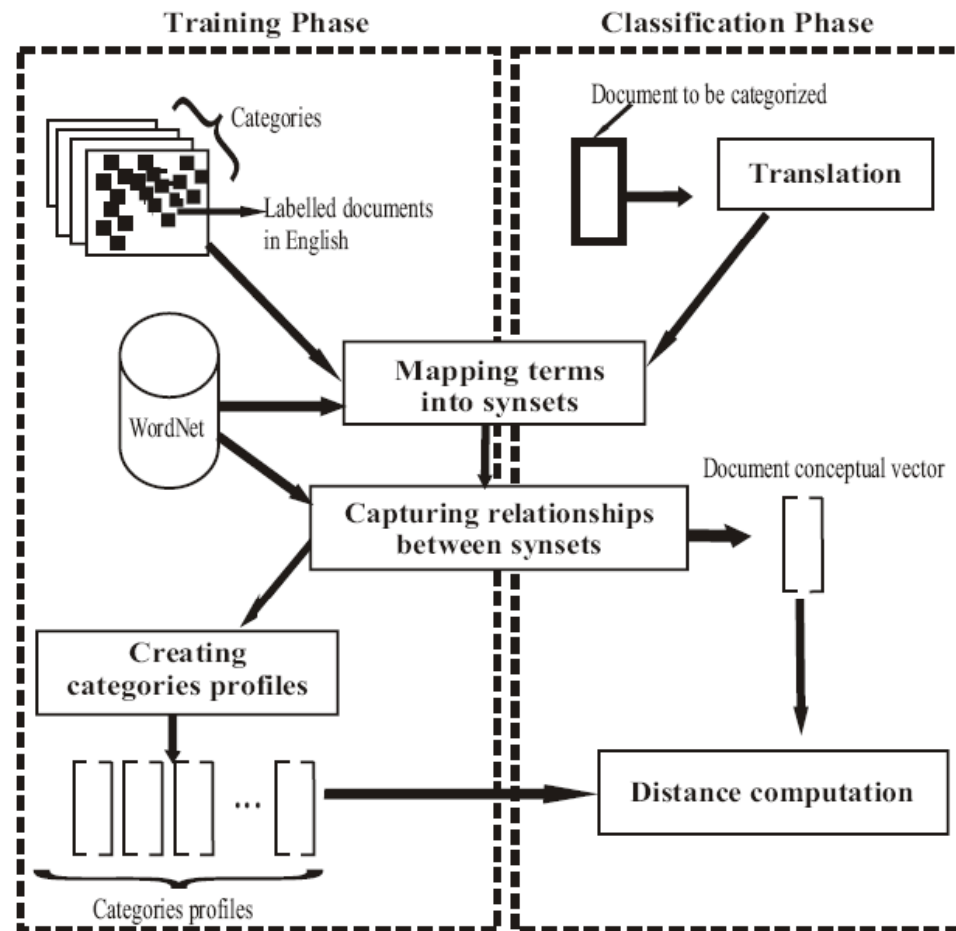
- CLIR: formulate a query in one language and retrieve documents in other languages
- CLTC: classify documents in several languages
- Both need similar linguistic resources

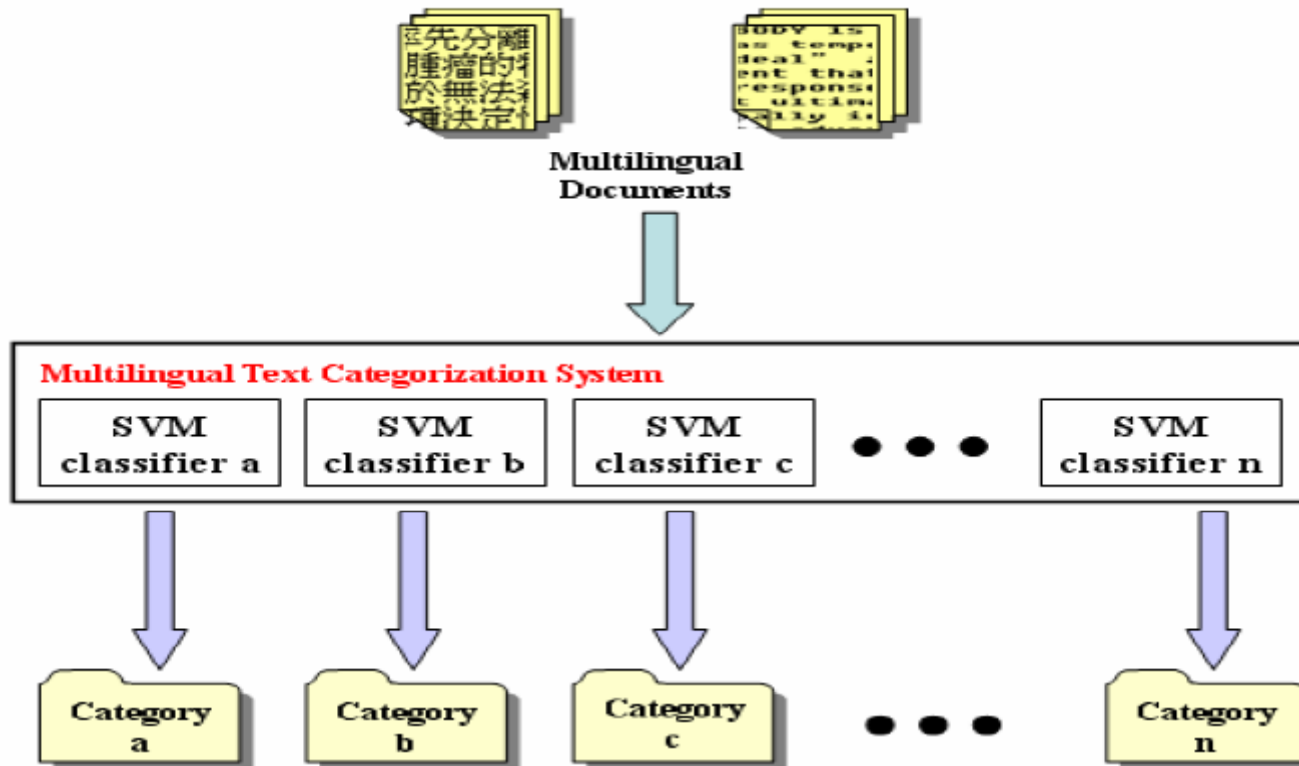
- CLTC need multilingual resources
 - some training documents in other languages
 - Non translated but supplied with a class label
 - Translated but without class label.
- ILO corpus: English and Spanish documents (only a couple of thousands each language)
- Reuter Corpus RCV2: about 500,000 reuters stories in 13 different languages
- Parallel corpora

- Poly-lingual
 - the system is trained using training examples from all the different languages.
 - a single classifier is build using a set of labeled training documents in all languages, which will classify documents from different languages.
 - no translation strategies

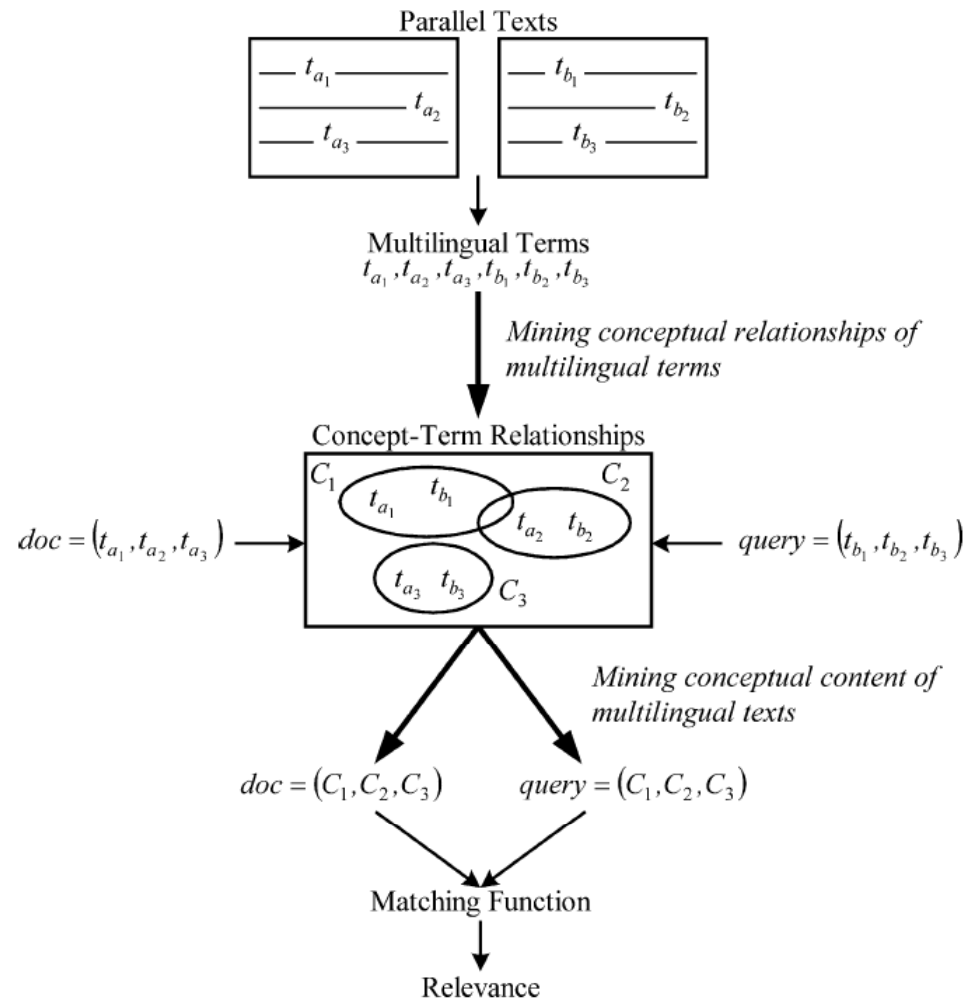
- Cross-lingual
 - use labeled training for one language to classify documents in other languages
 - Training-Set Translation: the labeled set is translated into the target language which then is used to train a classifier for this language.
 - Test-Set Translation: translate the unlabelled documents into one language (L1). To classify the unlabelled translated documents, the system is trained using the labelled training set for language (L1).

- Esperanto language:
 - use a universal reference language which all documents are translated to.
 - This universal language should contain all properties of the languages of interest and be organized in a semantic way.





CLTC example of architecture



- B. M. Amine, M. Mimoun, WordNet based Cross-Language Text Categorization Computer Systems and Applications, 2007. AICCSA apos;07. IEEE/ACS International Conference on Volume , Issue , 13-16 May 2007 Page(s):848 – 855
- R. Chau, C.H. Yeh, A multilingual text mining approach to web cross-lingual text retrieval, Knowledge-Based Systems 17(5/6) (2004) 219-27.
- W. S. Cooper, Inconsistencies and Misnomers in Probabilistic IR, in: A. Bookstein, Y. Chiaramella, G. Salton, V. V. Raghavan (Eds.), Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum), ACM, 1991, pp. 57-61.
- F. Damerau, T. Zhang, S. Weiss, N. Indurkha, Text Categorization for a Comprehensive Time-Dependent Benchmark, Information Processing and Management 40 (2) (2004) 209-221.
- G. M. Di Nunzio, A. Micarelli, Pushing "Underfitting" to the Limit: Learning in Bidimensional Text Categorization, in: R. L. de Mántaras, L. Saitta (Eds.), Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), including Prestigious Applicants of Intelligent Systems (PAIS 2004), IOS Press, 2004, pp. 465-469. G. M. Di Nunzio, Visualization and Classification of Documents: A New Probabilistic Model to Automated Text Classification, Bulletin of the IEEE Technical Committee on Digital Libraries (IEEE-TCDL) 2 (2).
- G. M. Di Nunzio, The Two-Dimensional Probabilistic Model for Automated Text Categorization, Ph.D. thesis, Department of Information Engineering, University of Padua, Padova, Italy (December 2005).
- G. Di Nunzio. Using Scatterplots to Understand and Improve Probabilistic Models for Text Categorization and Retrieval. International Journal of Approximate Reasoning, In Press 2009. <http://dx.doi.org/10.1016/j.ijar.2009.01.002>.
- S. Eyheramendy, D. D. Lewis, D. Madigan, On the Naive Bayes Model for Text Categorization, in: C. Bishop, B. Frey (Eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003.
- S. Eyheramendy, D. Madigan, A Novel Feature Selection Score for Text Categorization, in: Proceedings of International Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics (in conjunction with 2005 SIAM International Conference on Data Mining), 2005, pp. 1-8.
- M. Ferreira de Oliveria, H. Levkowitz, From Visual Data Exploration to Visual Data Mining: A Survey, IEEE Transactions on Visualization and Computer Graphics 9 (3) (2003) 378-394.
- M. Fuhr, Models in Information Retrieval, in: M. Agosti, F. Crestani, G. Pasi (Eds.), Lectures on Information Retrieval, Vol. 1980 of Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2000, pp. 21-50.
- A. Gliozzo, C. Strapparava, Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In Proceedings of the 21st international Conference on Computational Linguistics and the 44th Annual Meeting of the Association For Computational Linguistics (Sydney, Australia, July 17 - 18, 2006).
- M. A. Hearst, Modern Information Retrieval, Addison-Wesley (New York), 1999, Ch. User interfaces and visualization, pp. 257-323.
- T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, in: Nédellec and Rouveirol [30], pp. 137-142.



- D. Keim, Information Visualization and Visual Data Mining, IEEE Transactions on Visualization and Computer Graphics 8 (1) (2002) 1-8.
- C. Lee, H. Yang, S. Ma, A Novel Multilingual Text Categorization System using Latent Semantic Indexing. In Proceedings of the First international Conference on innovative Computing, information and Control - Volume 2 (August 30 - September 01, 2006). ICICIC.
- C-H. Lee, H-C. Yang, A Multilingual Text Mining Approach Based on Self-Organizing Maps. Appl. Intell. 18(3): 295-310 (2003)-
- D. Lewis, Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, in: Nedellec and Rouveirol [30], pp. 4-15.
- A. McCallum, K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification, in: AAI-98 Workshop on Learning for Text Categorization, AAAI Press, 1998, pp. 41-48.
- S. E. Middleton, N. R. Shadbolt, D. C. D. Roure, Ontological User Profiling in Recommender Systems, ACM Transaction on Information Systems 22 (1) (2004) 54-88.
- D. Mladenic, J. Brank, M. Grobelnik, N. Milic-Frayling, Feature Selection using Linear Classifier Weights: Interaction with Classification Models, in: J. Callan, N. Fuhr, W. Nejdl (Eds.), Peer-to-Peer Information Retrieval, ACM Press, New York, NY, USA, 2004, pp. 234-241.
- B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, B. Pearcy, C. Macdonell, J. Anvik, Visual Explanation of Evidence with Additive Classifiers, in: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI) and the 18th Innovative Applications of Artificial Intelligence (IAAI) Conference, July 16-20, 2006, Boston, Massachusetts, USA, AAAI Press, 2006.
- R. Rifkin, A. Klautau, In Defense of One-Vs-All Classification, Journal of Machine Learning Research 5 (2004) 101-141. 2004), including Prestigious Applicants of Intelligent Systems (PAIS 2004), IOS Press, 2004, pp. 465-469.
- S. E. Robertson, K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the American Society for Information Science 27 (1976) 129-146.
- P. Ruch, Query translation by text categorization. In Proceedings of the 20th international Conference on Computational Linguistics (Geneva, Switzerland, August 23 - 27, 2004).
- F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys 34 (1) (2002) 1-47.
- J. A. Walter, J. Ontrup, D. Wessling, H. Ritter, Interactive Visualization and Navigation in Large Data Collections using the Hyperbolic Space., in: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA, IEEE Computer Society, 2003, pp. 355-362.
- T. Zang, F. J. Oles, Text Categorization Based on Regularized Linear Classification Methods, Information Retrieval 4 (1) (2001) 5-31.





Outline

- Part I
 - Cross-lingual Text Categorization

- Part II
 - Log Analysis



- Query Classification
- Log Analysis
 - Web (HTTP) Log Analysis
 - Query (Search) Log Analysis
- Logging Digital Libraries
 - The European Library Case Study
- LogCLEF 2009



Query Classification

- Classification of user queries allows for increased effectiveness, and efficiency, in general-purpose Web search systems.
- Query classification poses a challenging problem, as web queries are very short.
- This data sparseness, and the dynamic nature of user queries makes the traditional methods of text categorization less effective.



- The problem of query classification can be split into two groups:
 - Query types
 - Query topics

- The term log is used today in the context of a systematic and chronological collection of events:
 - user log;
 - system log;
 - server log.
- We use the name "Web log" for the log file which contains the HTTP requests made from clients to the Web server of a Web site.
- We use the name "Query log" for the log file which contains the queries issued by users to search engines.



The importance of being logged...

- Logging is an implicit way of observing the user in an unbiased way
- Assuming adequate storage and infrastructure exists (not easy, see next), there are many limitations and biases to consider
 - Measure how and what, rather than “why”
 - Completely unlabeled
 - Can only measure the system being logged
 - Noisy
 - Don’t necessarily allow long-term studies



Managing logs

- Information contained in logs (and logs themselves) has to be managed efficiently for analyses.
- A database management system enable separation of the different entities recorded and facilitate data-mining and on-demand querying of the log data.
- Since log files usually come in a text format, there is the problem of how to transfer the huge amount of data into a DB.
- M. Agosti and G. M. D. Nunzio. Web log mining: A study of user sessions. 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries, June 2007.



- Several formats available for servers to build their log files.
- World Wide Web Consortium Extended Log File Format
 - It extends the Common Log File Format overthrowing its main limit, which was the fixed amount of information about each transaction;
 - control recorded data;
 - customize the choice of the fields of the record of a log to be stored.

- Web log file analysis began with the purpose to offer to Web site administrators a way to ensure adequate bandwidth and server capacity to their organization.
- It may offer advices about
 - a better way to improve the offer of Web content,
 - information about problems occurred to the users,
 - problems for the security of the site.



Web Log Analysis

- Web sites are created and adapted to make contents more easily accessible, using profiles found to make recommendations or to target users with ad hoc advertising.
- An ideal environment would dispose of exact history and information about a user, permitting to know his tastes and needs:
 - using user authentication,
 - or cookie files.
- Not always possible: privacy issues step in, and this kind of information may be not available.
- M. Agosti, G. Angelaki, T. Coppotelli, and G. M. D.Nunzio. Analysing http logs of a european DL initiative to maximize usage and usability. ICADL 2007 pages 35-44, 2007.



- Web log files have ordered fields to record activities:
 - *date*: Date, in the form of yyyy-mm-dd.
 - *time*: Time, in the form of hh:mm:ss.
 - *s-ip*: The IP of the server.
 - *cs-method*: The requested action. Usually GET for common users.
 - *cs-uri-stem*: The URI-Stem of the request.
 - *cs-uri-query*: The URI-Query, where requested.
 - *s-port*: The port of the server for the transaction.
 - *cs-username*: The username for identification of the user.
 - *c-ip*: The IP address of the client.
 - *cs(User-Agent)*: User-Agent of the Client. For a standard user this means the browser and other information about operative system.
 - *cs(Referer)*: The site where the link followed by the user was located.
 - *sc-status*: HTTP status of the request, that means the response of the server.
 - *sc-substatus*: The substatus error code.
 - *sc-win32-status*: The Windows status code.

Web Log File Example

- 2005-11-30 23:00:37 192.87.31.35 GET /index.htm - 80 - 152.xxx.xxx.xxx Mozilla/4.0+(compatible; ...
- 2005-11-30 23:00:38 192.87.31.35 GET /portal/index.htm - 80 - 152.xxx.xxx.xxx Mozilla/4.0+(...
- 2005-11-30 23:00:38 192.87.31.35 GET /portal/scripts/Hashtable.js - 80 - 152.xxx.xxx.xxx Mozilla/4.0+

- Gathering data from user agents
 - User agent string may offer great amount of information useful to optimize and personalize Web sites.
 - It is a valuable source that gives hints about
 - browsers,
 - operating systems used by users,
 - and even analyze the activity of crawlers.

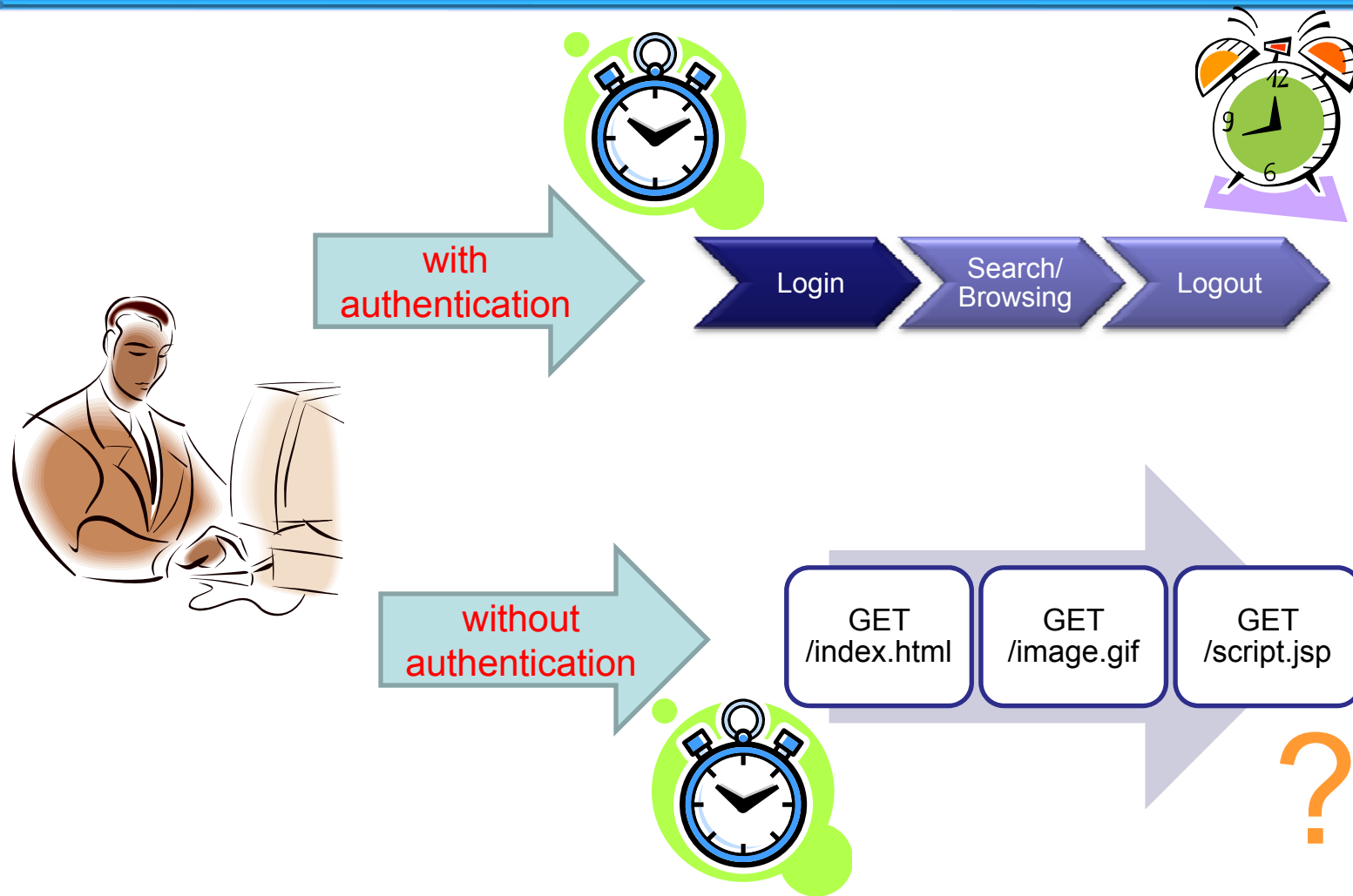
- There are no standards that define it.



Storing log files

- The design of the DataBase (DB) may begin with the analysis of the raw log files and followed by the identification of each entity with a top-down approach, starting from the main entity Request
- The main entities identified can be:
 - client;
 - useragent;
 - request;
 - uristem;
 - server.
- More entities can be identified to model post-hoc analyses on the identification of sessions according to different heuristics:
 - session;
 - heuristic.

Identification of sessions





Query Log Analysis

- The goal of a search engine is to have indexed the right pages for any user search, and to effectively retrieve them in response to the query supplied by the user.
- In order to accomplish these goals, the search engine must:
 - “Understand” the intent of the search query
 - Have indexed the correct pages, documents, or materials
 - Be able to effectively order and retrieve the right pages with respect to the intent of the user query.



Query Log Analysis

- Query logs are one of the largest sources of data potentially available to a search engine.
- The main advantage of query logs is that they have a diversity of tasks, queries, user experiences that is difficult, if not impossible, to duplicate in any other data source.
- However, this diversity is completely unlabeled:
 - We know a user did a query
 - We do NOT know what the user meant by it
 - We do NOT know whether the user was happy with the result.





Evaluation of Digital Libraries

- Digital libraries (DL) are complex systems for which specific set of important issues has to be considered, such as
 - the construct of evaluation
 - What was evaluated, what elements
 - the context of evaluation
 - Selection of a goal, a level of evaluation
 - the criteria
 - What parameters of performance were concentrate on
 - the methodology
 - What measures and measuring instruments were used





Logging Digital Libraries

- Transaction log analysis, a widely used technique that examines the activity of users in a given time session
 - collect and analyze all traces of a user's choices during a work session and thus create meaningful models of usage
 - New log formats?
- Need of research based techniques that require users' participation, observation and recording of their actions.
- M. Agosti. Log data in digital libraries. In M. Agosti, F. Esposito, and C. Thanos, editors, IRCDL, pages 115-122. DELOS: an Association for Digital Libraries, 2008.





Using different sources of logs

- The interaction between the user and a system can be analyzed
 - Explicitly, for example questionnaires,
 - Implicitly, for example log analysis
- A systematic use of the triangulation of different data collection techniques as a general approach can be used to get better knowledge of the information search process
- **M. Agosti, F. Crivellari, and G. M. Di Nunzio. A method for combining and analyzing implicit interaction data and explicit preferences of users. Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (ECIR 2009), April 2009.**





The European Library Case Study

- The European Library is a free service that offers access to the resources
 - of the 48 national libraries of Europe
 - in 20 languages
 - about 150 million entries across Europe
- TELplus is a project that is being taken forward by a consortium of 26 partners, national libraries and research centers to form a pool of research and innovation to provide value-adding services and products for TEL.





TEL Environment 1/2

The European Library - v1.6 - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti ?

http://www.theeuropeanlibrary.org/portal/index.html

Gmail - Inbox - airam...

La Biblioteca Europea Det europeaiske bibliotek La bibliothèque européenne De Europese Bibliotek Det europeiska biblioteket Avrupa Kütüphanesi Europska knjižnica Eiroopna Evropska knjižnica Az Európai Könyvtár Die Europäische Bibliothek Euroopan kirjasto Euroopa raamat

The European Library

HOME COLLECTIONS LIBRARIES ORGANISATION

Search Results History Favourites Help?

SEARCH

CHOOSE YOUR OWN COLLECTIONS:

- catalogues from national libraries
- digital collections
- audio and printed music
- images
- manuscripts
- maps
- journals
- children's literature
- theses and dissertations
- personal selection
- all collections from:

[Browse all the collections](#)

[Browse collections from:](#)

[Browse collections by subject](#)

[Search collections by description](#)

The "Historic Atlases of the Netherlands" Collection is [in maintenance](#)

The European Library searches the content of European national libraries.

NEWS 20/06/2008

First time visitors can now view The European Library [User Guide](#) in **18 languages**. The [User Guide](#) is designed to help new visitors discover all search possibilities offered by The European Library portal.

EXHIBITION [View Buildings](#)

Visit The European Library web-exhibition and view 300 pictures of the National libraries of Europe. Click [here](#) to enter the gallery.

TREASURES [View treasure](#)

[Missale Aboense National Library of Finland](#)

FAQ's | Jobs & Careers | Terms of service | Sitemap | Contact & Feedback | © 2005-2008 The European Library

The European Library is a project of the **CENL**

KB The European Library is hosted by the **Koninklijke Bibliotheek** - Koninklijke Bibliotheek, The Netherlands (KB)

June 17 2009, Pisa, Italy

Summer School





TEL Environment 2/2

The European Library 2.0 - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti 2

http://search.theeuropeanlibrary.org/portal/en/search/umberto+eco.query

DEI Webmail Gmail Yahoo Mail Facebook

La Biblioteca Europea Det europeiske bibliotek La bibliothèque européenne De Europese Bibliotheek
The European Library Det europeiska biblioteket Avrupa K tt phanesi Euroopa knjiznica Eiropas bibliot ka
na Evropska knjiznica Az Eur pai K nyvt r Die Europ ische Bibliothek Euroopan kirjasto Euroopa raamatuko

Language: English (eng) Register Login

HOME COLLECTIONS LIBRARIES ORGANISATION

Search Results History Favourites Help?

KatNUK: the catalogue of the Slovene National and University Library

Results 1 - 10 of 144 for 'umberto eco' (0.937 seconds.)

Print | jump to | GO

PREVIOUS NEXT

1 **Apokalipti ni in integrirani intelektualci : mnone komunikacije in teorije o mnoni kulturi**
Eco, Umberto
Type: text , analytic | Language: slv

Umberto Eco
Eco, Umberto
Type: text , analytic | Language: slv

La ricerca della lingua perfetta nella cultura europea
Eco, Umberto
Type: text , monograph | Language: ita

options

ABORT

search within results
 exclude from results
[Advanced search \(more options\)](#)
[Change the collections selection](#)

Matches for:
umberto eco

- SI KatNUK: the catalogue of the Slovene National and University Library
- DE Catalogue of the German National Library
- FI HELKA - Union catalogue of all Helsinki University libraries
- UK British Library integrated catalogue - Online catalogues of printed and electronic resources
- SK The Slovak Library - Catalogs and collections of Slovak libraries
- CH HELVETICAT : the catalogue of the Swiss National Library
- AT Online Catalogue of the Austrian National Library from 1992 onwards
- HU Amicue - Online catalogue of National Sz ch nyi Library of Hungary
- CZ Catalogue of the National Library of the

Type	text
monograph	
Author	Eco, Umberto
Title	La ricerca della lingua perfetta nella cultura europea
Date	1993
Subject	lingvistika - zgodovinski pregledi - filozofija jezika - evropska kultura - 15./19.st.
Edition	2a ed.
Language	ita
Publisher	Roma , Laterza
Format	VIII, 423 str. ilustr.
Extent	22 cm
Description	Bibliografija: str. 384-410
Is Part of	Fare l'Europa
ISBN	88-420-4287-0
UDC - subject	130.2(4)(091) 81:1(091)

Services

- AVAILABILITY at library
- LINK to other services
- COLLECTION HOMEPAGE

Options

- Save in favourites
- Send by E-mail

Quick links

Search all collections from:
Country

Browse collections from:
Country

The name of the rose = lme roze

June 17 2009, Pisa, Italy

Summer School





Different sources, different information (1)

- HTTP log data
- Reconstruct from the logs the history of the browsing activity
- Analysis
 - Traffic analysis/type of traffic
 - Session reconstruction/analysis
 - Geographic provenance





Different sources, different information (2)

- Action/search log data
- Reconstruct from the logs the history of queries and documents (i.e., search results) clicked on after the user has seen the result page
- Analysis
 - Session reconstruction/analysis
 - Rephrase queries
 - Switch to search material for a new topic





Different sources, different information (3)

- Questionnaire data
- Gather user preferences and satisfaction
- Analysis
 - Parts of the Web site
 - Clarity of the Web site
 - Level of satisfaction



User Studies

- A user's study was conducted by UNIPD (November/December 2007, April/May 2008)
- Different groups of students
 - Humanities: 188
 - Statistics: 6
 - Psychology: 22
- Total of 216 students
 - Only 155 were correctly found in the logs





Towards LogCLEF 2009

- Log analysis initiatives in CLEF
- iCLEF (since 2001)
 - Investigate cross-language searching as an interactive task, examining the process as well as the outcome.
- LogCLEF (pilot task 2009)
 - <http://www.uni-hildesheim.de/logclef/>
 - analysis of queries as expression of user behavior
 - Log Analysis and Geographic Query Identification (LAGI)
 - Log Analysis for Digital Societies (LADS)



- M. Agosti, editor. Information access through search engines and digital libraries. Springer, Berlin, Germany, 2008.
- M. Agosti. Log data in digital libraries. In M. Agosti, F. Esposito, and C. Thanos, editors, IRCDL, pages 115-122. DELOS: an Association for Digital Libraries, 2008.
- M. Agosti, G. Angelaki, T. Coppotelli, and G. M. D. Nunzio. Analysing http logs of a european DL initiative to maximize usage and usability. ICADL 2007 pages 35-44, 2007.
- M. Agosti, F. Crivellari, and G. M. Di Nunzio. A method for combining and analyzing implicit interaction data and explicit preferences of users. Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (ECIR 2009), April 2009.
- M. Agosti and G. M. D. Nunzio. Web log mining: A study of user sessions. 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries, June 2007.
- P. G. Anick. Using terminological feedback for web search refinement: a log-based study. In SIGIR, pages 88-95. ACM, 2003.
- S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. A. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In M. Sanderson, K. J  arvelin, J. Allan, and P. Bruza, editors, SIGIR, pages 321-328. ACM, 2004.
- N. Buzikashvili. An exploratory web log study of multitasking. In Efthimiadis et al. [12], pages 623-624. [7] N. Buzikashvili. Sliding window technique for the web log analysis. In Williamson et al. [32], pages 1213-1214.
- L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, and M. Dahlin, editors. Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006. ACM, 2006.
- J. Chen and T. Cook. Mining contiguous sequential patterns from web logs. In Williamson et al. [32], pages 1177-1178.
- S.-L. Chuang, H.-T. Pu, W.-H. Lu, and L.-F. Chien. Auto-construction of a live thesaurus from search term logs for interactive web search. In SIGIR, pages 334-336, 2000.
- H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In WWW 2002, pages 325-332, 2002.
- E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. J  arvelin, editors. SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 2006. ACM, 2006.
- N. Fuhr, P. Hansen, M. Mabe, A. Micsik, and I. Solvberg. Digital libraries: A generic classification and evaluation scheme. In P. Constantopoulos and I. Solvberg, editors, ECDL, volume 2163 of Lecture Notes in Computer Science, pages 187-199. Springer, 2001.
- N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovas, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Solvberg. Evaluation of digital libraries. International Journal on Digital Libraries (IJDL), 8(1):2738, 2007.
- W. Gao, C. Niu, J.-Y. Nie, M. Zhou, J. Hu, K.-F. Wong, and H.-W. Hon. Cross-lingual query suggestion using query logs of different languages. In Kraaij et al. [19], pages 463-470.
- M. A. Goncalves, G. Panchanathan, U. Ravindranathan, A. Krowne, E. A. Fox, F. Jagodzinski, and L. N. Cassel. The xml log standard for digital libraries: Analysis, evolution, and deployment. In JCDL, pages 312-314. IEEE Computer Society, 2003.



- C. Grimes, D. Tang, and D. M. Russell. Query logs alone are not enough. In E. Amitay and C. G. M. J. Teevan, editors, Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007), May 2007.
- R. Hu, W. Chen, P. Bai, Y. Lu, Z. Chen, and Q. Yang. Web query translation via web log mining. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, SIGIR, pages 749-750. ACM, 2008.
- P. Ingwersen and K. Järvelin. The Turn. Springer, The Netherlands, 2005. [18] T. Koch, A. Ardão, and K. Golub. Browsing and searching behavior in the renardus web service a study based on log analysis. In H. Chen, H. D. Wactlar, C. chih Chen, E.-P. Lim, and M. G. Christel, editors, JCDL, page 378. ACM, 2004.
- W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors. SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007. ACM, 2007.
- M. Maslov, A. Golovko, I. Segalovich, and P. Braslavski. Extracting news-related queries from web query log. In Carr et al. [8], pages 931-932.
- J. C. Miller, G. Rae, and F. Schaefer. Modifications of kleinberg's hits algorithm using matrix exponentiation and weblog records. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, SIGIR, pages 444-445. ACM, 2001.
- J. Parikh and S. Kapur. Unity: relevance feedback using user query logs. In Efthimiadis et al. [12], pages 689-690.
- N. Pharo and K. Järvelin. The SST method: a tool for analysing Web information search processes. Information Processing & Management, 40(4):633-654, July 2004.
- S. Sekine and H. Suzuki. Acquiring ontological knowledge from query logs. In Williamson et al. [32], pages 1223-1224.
- X. Shi and C. C. Yang. Mining related queries from search engine query logs. In Carr et al. [8], pages 943-944.
- R. Srikant and Y. Yang. Mining web logs to improve website organization. In WWW 2001, pages 430-437, 2001.
- Y. Sun, K. Xie, N. Liu, S. Yan, B. Zhang, and Z. Chen. Causal relation of queries from temporal logs. In Williamson et al. [32], pages 1141-1142.
- J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. History repeats itself: repeat queries in yahoo's logs. In Efthimiadis et al. [12], pages 703-704.
- J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In Kraaij et al. [19], pages 151-158.
- X. Wang and C. Zhai. Learn from web search logs to organize search results. In Kraaij et al. [19], pages 87-94.
- R. W. White, C. L. A. Clarke, and S. Cucerzan. Comparing query logs and pseudo-relevance feedback for web-search query refinement. In Kraaij et al. [19], pages 831-832.
- C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors. Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, 2007. ACM, 2007.
- Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In Carr et al. [8], pages 1039-1040.

