Database and Information Systems

11. Data Warehouses and OLAP
12. Data Mining
13. Semi-structured Data
14. Document Retrieval
15. Web Mining
16. Content Extraction
17. Multimedia data

Document Retrieval

Readings:

- Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, Addison Wesley, Harlow etc. 1999, chapter 2 and 3.

Document Retrieval

- goal: content-based access to unstructured data
- primary application areas
  - electronic libraries
  - search engines (WWW)
  - knowledge management
- result: (gradual) estimation of the relevance of a given document with respect to a query

Document Retrieval

- Document Representation
- Preprocessing
- Retrieval Models
- Quality Measures
- Text Mining

Document Representation

- selection of relevant documents based on keywords
- descriptors: keywords about the text
  - free choice of descriptors
  - restricted choice of descriptors (thesaurus)
- index terms: keywords from the text
- manual vs. automatic selection of keywords

Document Representation

- keyword k_i (index terms): lexical token used to characterise a document
- weights w_{ij}: numerical value describing the "importance" of a term k_i for the content of the document d_j

$$w_{ij} \begin{cases} > & 0 \quad \forall k_i \in d_j \\ = & 0 \quad \text{else} \end{cases}$$

- each document d_j is characterised by a vector of weights $\vec{d}_i$
- usually many keywords
  → vectors of extremely high dimensionality

Document Retrieval

- Document Representation
- Preprocessing
- Retrieval Models
- Quality Measures
- Text Mining

Preprocessing

- removal of formatting information
- removal of non-textual information
  - graphics, numbers, ...
- removal of stop words (function words)
  - high frequency words
  - assumption: contribute little to the content of a text
  - English: *the, a, an, and, or, of, for, ...*
  - German: *der, die, das, den, dem, und, oder, ...*
- removal of interpunction characters
- removal of infrequent word (too many)

Preprocessing

- normalization of word forms
- lemmatization: determining the (canonical) citation form
    - removal of inflectional endings
    - English:
      cars → *car*, *glasses* → *glass*,
      children → *child*, *men* → *man*
    - German:
      Bilder → *Bild*, *Maler* → *Maler*,
      Bildern → *Bild*, *Malern* → *Maler*,
      Zeiten → *Zeit*, *Fallen* → *Falle*,
      Äste → *Ast*, *Ähren* → *Ähre*
    - splitting of compounds:
      Weltuntergang → *Welt* + *Untergang*
      Innovationsschwäche → *Innovation* + *Schwäche*

Document Retrieval

- Document Representation
- Preprocessing
- Retrieval Models
- Text Mining

Retrieval Models

- Boolean Model
- Vector Space Model
- Weighted Boolean Model
- Probabilistic Model
- Latent Semantic Indexing

Boolean Model

- document description: $w_{ij} \in \{0, 1\}$
- query is a logical formula on index terms
  - binary weights
  - only full match
  - no grading of importance
  - no ranking of results

Boolean Model

- specification of descriptors
  ```
  ['databases'] in Text
  ```
- specification of a set of descriptors
  ```
  ['databases', 'information systems'] in Text
  ```
- use of logical connectors
  ```
  ['databases' AND 'multimedia'] in Text
  ```
- proximity queries: context sensitive patterns
  ```
  ['object-relational' WORD(-1) 'databases']
        in Text
  ['object-oriented' SAME_SENTENCE 'databases']
        in Text
  ['object-oriented' PARAGRAPH(2) 'databases']
        in Text
  ```
- weighted queries: how important is a descriptor
  ```
     hotel:0.8 AND seaside:0.5 AND view:0.2
  ```

Vector Space Model

- document and query are represented as a vector of weights $\vec{q}$
- degree of matching between two vectors is defined as a similarity function: e.g. cosine of two vectors

$$\begin{aligned} sim(d_i, q) &= \frac{\vec{d_i} \cdot \vec{q}}{|\vec{d_i}| \cdot |\vec{q}|} \\ &= \frac{\sum_{i=1}^{t} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \end{aligned}$$

  - partial match is possible
  - results can be ranked according to decreasing similarity

Vector Space Model

- Which measure to use as term weights w_{ij}?
- tf: normalized frequency of a term k_i in a document d_j

$$tf_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

f_{ij} frequency of term k_i in document d_j

- idf: inverse document frequency of a term k_j

$$idf_i = \log \frac{N}{|\{d_j | k_i \in d_j\}|}$$

N number of documents

Vector Space Model

- tf-idf: combining tf and idf

$$w_{ij} = tf_{ij} \cdot idf_i = \frac{f_{ij}}{\max_k f_{kj}} \cdot \log \frac{|\{d_1, ..., d_N\}|}{|\{d_j | k_i \in d_j\}|}$$

- the higher the frequency of a term in a document the better this term reflects the content of the document
- the smaller the set of documents in which a term occurs the better is the utility of the term to describe the content of a document (i.e. words occuring in every text are ignored)

- simple, fast and reliable

Vector Space Model

- retrieval of similar documents
- relevance feedback: "Give me more of this"
  - using a sample documents as extended query
  - more general: classify the retrieved documents as being useful (D_+) or not (D_-)
  - compute a new description vector as the weighted sum of the original query and the classified documents

$$\vec{q}\,' = \alpha\, \vec{q} + \beta \sum_{D_+} \vec{d} - \gamma \sum_{D_-} \vec{d}$$

- query as a generalised document: "Ask Jeeves about"

Weighted Boolean Model

- terms in the document are weighted (e.g. tf-idf)
- terms in the query not
- special operators for the logical connectors
  e.g. P-norm model for $p = 2$

$$sim(\vec{d}_j, \vec{q}_{and}) = 1 - \sqrt{\frac{\sum_{i=1}^{n} (1 - w_{ij})^2}{2}}$$

$$sim(\vec{d}_j, \vec{q}_{or}) = 1 - \sqrt{\frac{\sum_{i=1}^{n} w_{ij}^2}{2}}$$

n: number of terms in the query

Probabilistic Model

- trying to maximize the probability that a set of documents is relevant to the user
- user feedback can be used to adapt the model
- similarity

  $$sim(d_j, q) = \frac{p(R|\vec{d_j})}{p(\bar{R}|\vec{d_j})}$$

  R: set of documents considered to be relevant
  $\bar{R}$: complement of R
  weights are binary: $w_{ij} \in \{0, 1\}$
- probabilities are query dependent!

Probabilistic Model

- using Bayes' rule

$$sim(d_j, q) = \frac{p(\vec{d_j}|R) \cdot p(R)}{p(\vec{d_j}|\bar{R}) \cdot p(\bar{R})}$$

- $p(R)$ and $p(\bar{R})$ are the same for all documents, i.e. they do not influence a document selection based on maximum probability

$$sim(d_j, q) \sim \frac{p(\vec{d_j}|R)}{p(\vec{d_j}|\bar{R})}$$

Probabilistic Model

- assuming independence between index terms

$$sim(d_j, q) \sim \frac{\prod_{i, w_{ij}=1} p(k_i|R) \cdot \prod_{i, w_{ij}=0} p(\bar{k}_i|R)}{\prod_{i, w_{ij}=1} p(k_i|\bar{R}) \cdot \prod_{i, w_{ij}=0} p(\bar{k}_i|\bar{R})}$$

- taking logarithm, ignoring constant factors, and using the law of overall probability

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{iq} \cdot w_{ij} \cdot \left(\log \frac{p(k_i|R)}{1 - p(k_i|R)} + \log \frac{1 - p(k_i|\bar{R})}{p(k_i|\bar{R})} \right)$$

- Where do the probabilities $p(k_i|R)$ and $p(k_i|\bar{R})$ come from?

Probabilistic Model

- recursive procedure (expectation maximisation):
  - assume initial values for $p(k_i|R)$ and $p(k_i|\bar{R})$
  - if no other information available:

    $$p(k_i|R) = \text{ const,} \qquad \text{i.e. } p(k_i|R) = 0.5$$

    $$p(k_i|\bar{R}) = \frac{|D_i|}{|D|}$$

    D: set of all documents; $D_i = \{d_j \in D | k_i \in d_j\}$
  - retrieve a document set

Probabilistic Model

- recursive procedure (cont.)
  - recompute $p(k_i|R)$ on the retrieved documents V assuming all retrieved documents are relevant

$$p(k_i|R) = \frac{|V_i|}{|V|}, \quad V_i = \{d_j \in V | k_i \in d_j\}$$

  - recompute $p(k_i|\bar{R})$ assuming all not-retrieved documents are irrelevant

$$p(k_i|\bar{R}) = \frac{|D_i| - |V_i|}{|D| - |V|}$$

  - or: use user relevance feedback to improve the estimate
  - continue until a termination criterion is met

Latent Semantic Indexing

- keyword-based retrieval can give poor results because of
  - ambiguity: a keyword can describe quite different concepts

    tree → plant, structure
  - variety: more than a single keyword can be used to describe a concept

    program, software, code →
    piece of text written in a programming language
- document retrieval should be based on concepts rather than keywords
- idea: concepts are characterised by common contexts in which they appear

Latent Semantic Indexing

- term-document association matrix: M ($t \times N$-matrix)
- each matrix element is the weight of the term (row) in a document (column)
- M is decomposed into three components

$$M = K \cdot S \cdot D^T$$

- K: matrix of eigenvectors derived from the quadratic matrix $M \cdot M^T$ (describes terms as vectors of derived orthogonal factors ("concepts"))
- D^T: matrix of eigenvectors derived from the quadratic matrix $M^T \cdot M$ (describes documents as vectors of derived orthogonal factors)
- S: $r \times r$ diagonal matrix of singular (scaling) values ($r = \min(t, N)$)

Latent Semantic Indexing

- keep only the s largest singular values from S together with the corresponding columns in K and D^T
    - dimensionality reduction
- M_s: matrix of rang s which approximates M with minimum square error

  $$M_s = K_s \cdot S_s \cdot D_s^T \approx M$$

  s: dimensionality of the reduced concept space $(s < r)$
- s should be
    - large enough to accomodate all the structure in the data
    - small enough to suppress all the non-relevant details

Latent Semantic Indexing

- relationships between two documents can be obtained from

$$\begin{aligned} M_s^T \cdot M_s &= (K_s \cdot S_s \cdot D_s^T)^T \cdot K_s \cdot S_s \cdot D_s^T \\ &= D_s \cdot S_s \cdot K_s^T \cdot K_s \cdot S_s \cdot D_s^T \\ &= D_s \cdot S_s \cdot S_s \cdot D_s^T \\ &= (D_s \cdot S_s) \cdot (D_s \cdot S_s)^T \end{aligned}$$

- $sim(d_i, d_j) = e_{ij}(M_s^T \cdot M_s)$
- model the query as a pseudo document: the corresponding row in $M_s^T \cdot M_s$ contains the ranks of all documents with respect to this query

Latent Semantic Indexing

- $s \ll t, s \ll N$
  - efficient indexing scheme
  - elimination of noise from the data
  - removal of redundancy
- problem: open document collections

Document Retrieval

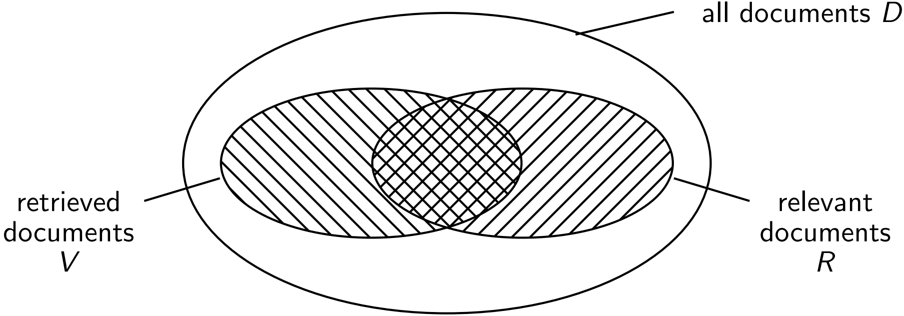
- other similarity-based models
    - fuzzy set model
    - generalised vector space model
    - neural networks
    - bayesian networks
    - inference network model
    - belief network models
    - ...

Document Retrieval

- Document Representation
- Preprocessing
- Retrieval Models
- Quality Measures
- Text Mining

Quality Measures

- recall: fraction of the relevant documents which has been retrieved
- precision: fraction of the retrieved documents which is relevant



Quality Measures

- precision:

$$precision = \frac{|V \cap R|}{|V|}$$

- recall:

$$recall = \frac{|V \cap R|}{|R|}$$

- fallout:

$$fallout = \frac{|R - V|}{|D - R|}$$

Quality Measures

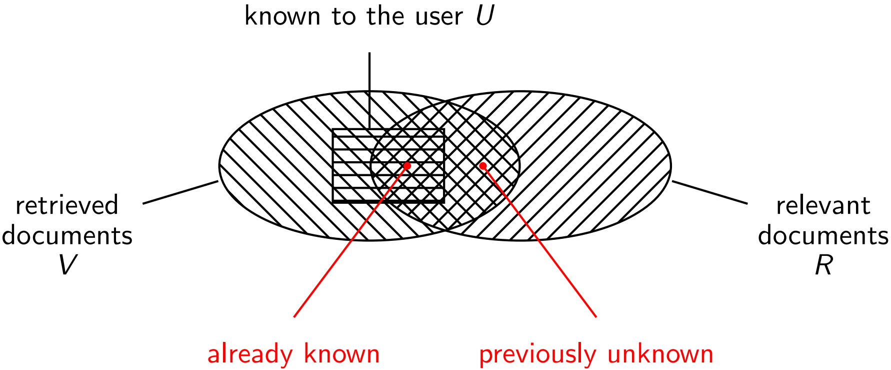
- precision and recall are antagonistic measures:
  - maximizing recall reduces precision
  - maximising precision reduces recall
- if ranked documents are inspected incrementally, both figures vary over time
- open document collections: recall cannot be computed
  Which are the relevant documents in the web?

Quality Measures

- relevance is
  - a subjective notion
  - specific to a single query
- quality measures for test suites of queries: weighted sum over precisions at a given recall level

Quality Measures

- user-oriented measures: set of relevant documents is unknown



known to the user U

retrieved documents V

relevant documents R

already known

previously unknown

Quality Measures

- coverage: fraction of documents which is known to the user and has been retrieved

$$coverage = \frac{|U \cap V|}{|U|}$$

- novelty: fraction of the retrieved and relevant documents previously unknown to the user

$$novelty = \frac{|(R \cap V) - U|}{|R \cap V|}$$

Quality Measures

- relative recall: found relevant documents divided by expected relevant documents
- recall effort: expected documents divided by documents to be read

Quality Measures

- objective evaluation methods
  - standardised retrieval tasks on manually prepared text corpora (”gold standard”)
  - text retrival conferences (TREC)
  - annual competition organized by the National Institute of Standards (NIST)

Document Retrieval

- Document Representation
- Preprocessing
- Retrieval Models
- Quality Measures
- Text Mining

Text Mining

- Classification
- Clustering

Classification

- supervised learning: class assignment for the training data is given
- applications:
  - routing of emails
  - filtering and grouping of online news
- train a neural network or a probabilistic classifier

Clustering

- unsupervised learning: only maximum number of classes is given
- applications:
  - presentation of retrieval results
  - navigation tools for text collections
- class coherence: term frequency (tf)
- class distinction: inverse document frequency (idf)
- weighting of descriptors: $tf \cdot idf$

Evaluation

- example: Excite (Frankfurter Rundschau online)
  - ranking
  - retrieval of similar documents
  - clustering

Evaluation

- query: ”schmutzige Bombe”
- ranking: 20 documents, 5 relevant:

70% Zugang zu radioaktivem Material
68% Bedrohung durch radioaktive Bomben
64% Anschläge auf US-Marine geplant
54% Bombe aus 2. Weltkrieg entschärft
52% Bombenanschlag in Israel
50% Bedrohung durch radioaktive Bomben
50% Stadt ist zu schmutzig
48% Atomkraft ist schmutzige Energie
48% schmutzige Strassen
48% Bombenanschlag auf israelisches Tanklager
44% Islamisten-Prozess, tödliche Bombe
44% Beziehungsdrama, Bombenattrappe entdeckt
44% Bombenanschlag in Israel
44% Entschädigung für libyschen Bombenanschlag
44% Bedrohung durch radioaktive Bomben
42% Bombenanschlag in Israel
42% Dr. Seltsam: Wie sieht eine Atombombe aus?
42% schmutziges Grün auf Ausstellung
42% Bedrohung durch radioaktive Bomben
42% Bomben im 2. Weltkrieg

Evaluation

- clustering
- group 1: bush, lämmle, palästinenser, gray, bombe
  9 documents, 3 relevant
  unkontrollierter Zugang zu radioaktivem Material, Bedrohung durch radioaktive Bomben, Anschläge auf US-Marine geplant, Bombe aus 2. Weltkrieg entschärft, Bombenanschlag in Israel, Bedrohung durch radioaktive Bomben, Bombenanschlag in Israel, Bombenanschlag in Israel, Militärschlag gegen Irak
- group 2: ich, jeroen, willems, monolog, sehr
  5 documents, 0 relevant
  schmutzige Servietten, Umgang mit technischen Produkten, Bombenangriffe auf Serbien, moralische Kategorien, schmutzige Wäsche

Evaluation

- group 3: kpnqwest, boston, kirche, priester, massachusetts
  5 documents, 0 relevant
  Entschädigung für libyschen Bombenanschlag, nicht schmutzig machen, Druckwelle einer Bombe, schmutzige Details, Bomben gegen das Internet
- group 4: kensington, palace, leigh, königin, apartments
  1 documents, 0 relevant
  schmutzige Schuhe, schmutzige Schuhe
- group 5: kanata, ottawa, wäsche, ottawas, wäscheleinen
  2 documents, 0 relevant
  schmutzige Wäsche waschen, schmutzige Wäsche waschen

Evaluation

- group 6: fifa, blatter, hayatou, addo, seenot
  2 documents, 0 relevant
  schmutzige Tricks, schmutziges Spiel
- group 7: stadionbad, sprungturm, wasserspringer, beckenrand, wasser
  1 document, 0, relevant
  platschende Bombe
- group 8: darmkrebs, riemann, darmkrebse, darmkrebsrisiko, früherkennung
  1 document, 0 relevant
  Dunstkreis des Schmutzigen
- group 9: jerome, ungebetenen, bronx, audrah, karst
  1 document, 0 relevant
  Bombenattentat im Kriminalroman

Evaluation

- separaration of ambiguous words (1): Strom
- group 1: biblis, edf, kwk, ovag, block

  ... Attacke gegen den staatlichen französischen Energiekonzern EdF, da nicht zu erkennen sei, woher der Strom stamme.

  ... ein Gesetz zur Förderung von Strom aus umweltfreundlichen Kraft-Wärme-Kopplungs-Anlagen (KWK) auf den Weg gebracht.

  ... Oberhessische Versorgungsbetriebe AG (Ovag) kann sich über ihr Strom-Geschäft nicht beklagen.

  Bundeswirtschaftsminister Werner Müller droht dem französischen Stromkonzern Electricité de France (EdF) mit einem Importboykott.

  ... wie Motorroller, Leichtmofas, Power-Bikes und E-Kickboards, die ihre Leistung aus rein regenerativen Stromquellen beziehen.

  ... hatte es allein in Rheinland-Pfalz mehr als 60 öffentliche und private Badeanstalten an dem Strom gegeben.

  ... zwei Boote der Vereinten Nationen auf dem Kongo eingetroffen ... Bis vor kurzem war der Strom Kampfschauplatz gewesen.

Evaluation

- separation of ambiguous words (2): Gewinn
- group 1: dm, kl, lotto, gewinnt, dmez

  Glücksspiel: 7 documents, 7 fitting

  Vor einigen Jahren hatte sich Frauchen beim Gassigehen die Hausnummern gemerkt, an denen Bodo das Bein hob, entsprechende Kreuze im Lotto-Schein gesetzt und damit einen Volltreffer gelandet: "Bodo hat 1,9 Millionen Mark erpinkelt", bilanziert Hans-Joachim Schmitz von der Lotto-Gesellschaft Rheinland-Pfalz. Gassigehen mit Bodo könnte sich derzeit besonders lohnen: 36 Millionen Mark (18,4 Millionen Euro) winken im Samstag-Lotto, nachdem der Jackpot am Mittwoch wieder nicht geknackt wurde, sagte Schmitz.

  Geknackt ist er, der Lotto-Jackpot von 36 Millionen Mark, in den die Verlierer der wöchentlichen Hoffnungszulage einmal mehr duldsam eingezahlt hatten, um wieder leer auszugehen. Irgendwo gibt es zwei Gewinnende im Lande, die nun auf ihren Schein starren, um sich zu vergewissern, dass sie sich nicht verzählt oder die Superzahl mit der Endnummer des Spiels 77 verwechselt haben. Für alle anderen gilt, dass Lotto die Illusion auf einen finalen Kurswechsel eines Spiels ist, das letztlich doch wieder bloß weiter geht...

Evaluation

- group 2: audi, infomatec, lufthansa, harlos, piloten
  finanzieller Gewinn: 8 documents, 6 fitting
- group 3: eintracht, nils, schuss, 2, bhc
  sportlicher Gewinn: 5 documents, 4 fitting
- group 4: indus, kill, kapitalismus, kobank, risikostreuung
  finanzieller Gewinn: 5 Dokumente, 5 fitting
- group 5: tour, de, france, armstrong, etappe
  sportlicher Gewinn: 4 documents, 4 fitting

Evaluation

- group 6: edf, orlando, mafia, fiat, bandenmäßig

  diverse Gewinne:

  große Mengen an Drogen ..., um sie mit Gewinn ... weiterzuverkaufen.
  ... auch drei Geldhäuser für ihr geplantes Joint Venture gewinnen.
  ... der Gewinn aller Direktmandate .

- group 7: a, bildungslücke, 63, box, magazin

  Preisausschreiben: 3 documents, 3 fitting

- group 8: incognegro, uh, oh, funkiness, hiphop

  ... Gänseblümchen gewinnen .

Evaluation

- group 9: olympischen, olympia, olympiabewerbung, machbarkeitsstudie, ioc
  ... innerdeutsches Rennen ... gewinnen ....
- group 10: dosenpfand, anhänger, umfrage, befürworten
  Mehrheit dafür gewinnen ...
- group 11: caisse, ware, visier, aufsicht, mit, bankenriese
  finanzieller Gewinn
- group 12: misslicher, scharon, palästinenser, arafats
  politischer Gewinn