Database and Information Systems

11. Data Warehouses and OLAP
12. Data Mining
13. Semi-structured Data
14. Document Retrieval
15. Web Mining
16. Text Mining
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
- Retrieval Models
- Quality Measures
- Document Management

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

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

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

Document Retrieval

- Document Representation
- Preprocessing
- Retrieval Models
- Document Management

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_i

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

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 document 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

Document Retrieval

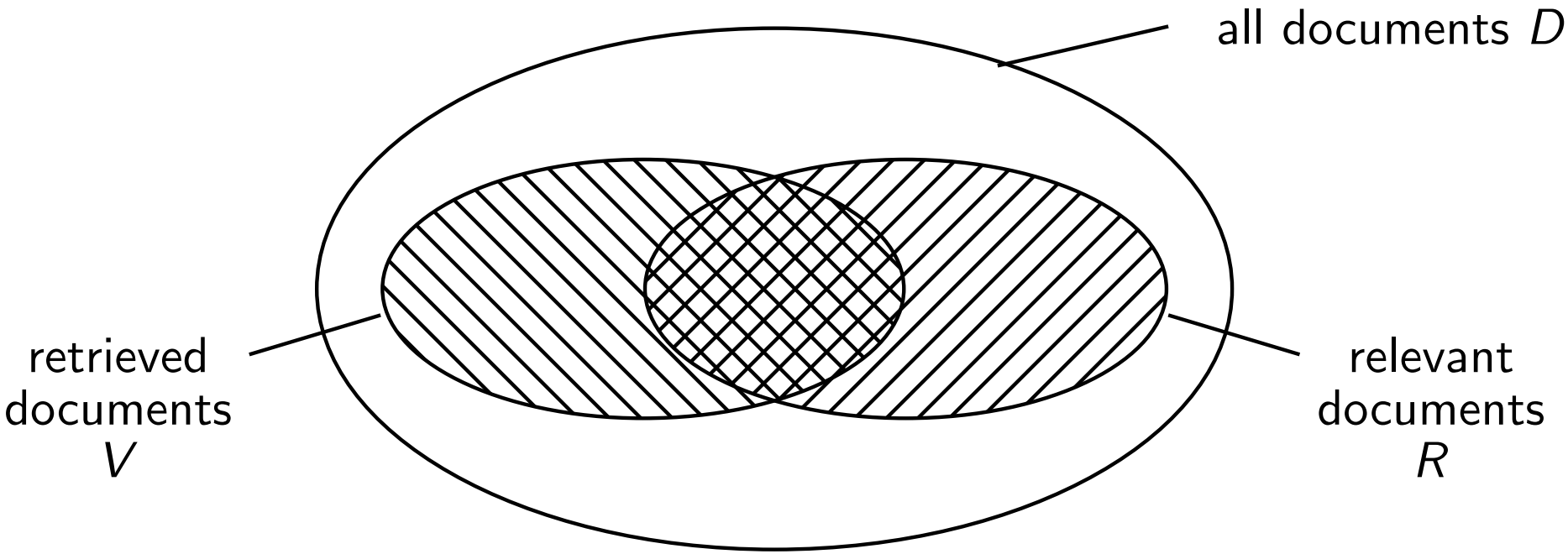
- other similarity-based models
  - fuzzy set model
  - generalised vector space model
  - neural networks
  - bayesian networks
  - inference network model
  - belief network models
  - ...

Document Retrieval

- Document Representation
- Retrieval Models
- Quality Measures
- Document Management

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

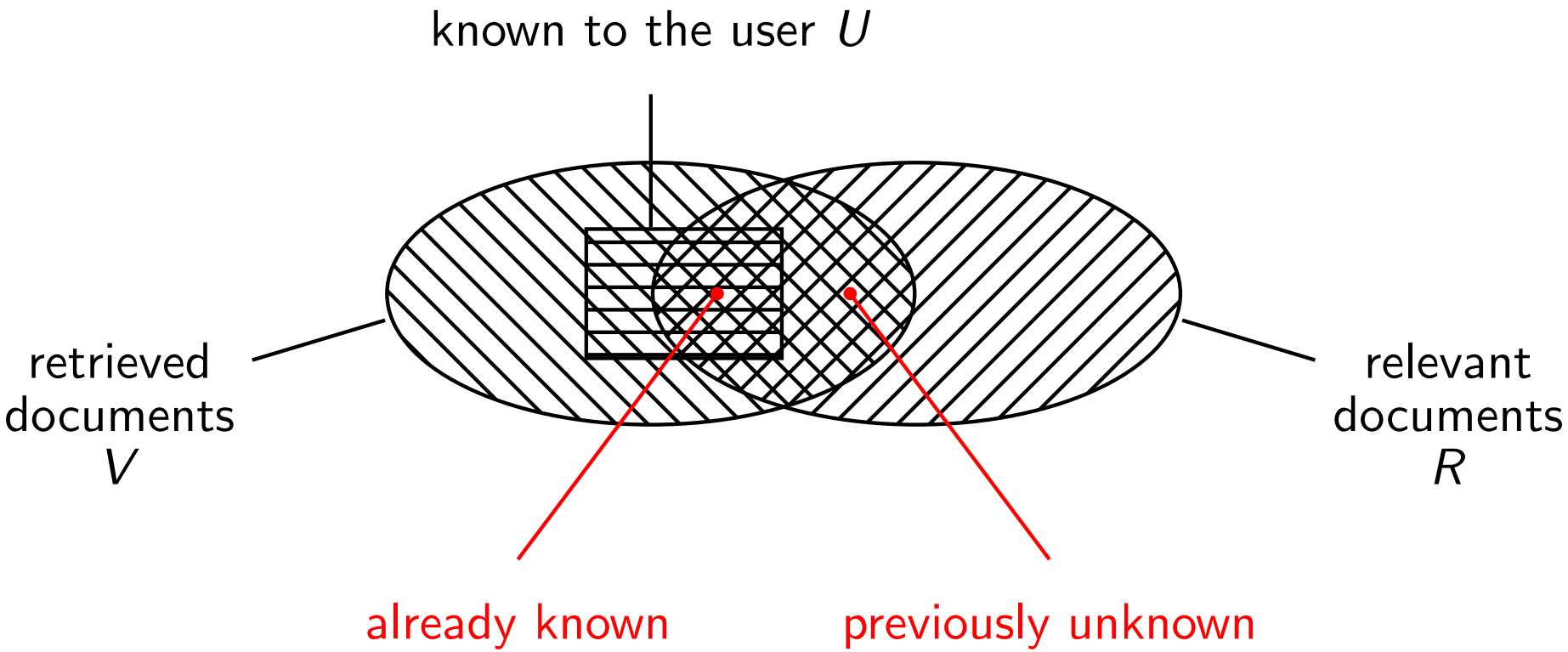
- precision and recall are antagonistic measures:
  - maximizing recall reduces precision
  - maximising precision reduces recall
- if ranked documents are inspected incrementally, both figures vary over time
- open document collections: recall cannot be computed
  Which are the relevant documents on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



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
- Retrieval Models
- Quality Measures
- Document Management

Document Management

- 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

Database and Information Systems

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

Readings

- Soumen Chakrabarti: Mining the Web. Morgan Kaufmann Publishers, Amsterdam etc. 2003

Web Mining

- Markup Analysis
- Link Context Analysis
- Link Structure Analysis
- Page Role Classification
- Problems of Link Analysis

Markup Analysis

- web documents contain more information than pure text
  - metainformation
    - description
    - keywords
  - document structure
    - URI
    - title
    - headlines
    - font tags (strong, bold, emphasize)
    - text body
    - ...
  - different kinds of information can be weighted differently

Markup Analysis

- markup is used for spamming: "invisible" text
  - metainformation
  - color
- link frequency information is also not reliable
  - web pages are mirrored
  - removal of duplicates becomes necessary
- evaluating the quality of a page:
  - detailed analysis of the link structure

Link Context Analysis

- the web is more than a collection of documents and links
- links are *embedded* in a web-page
- anchor text:

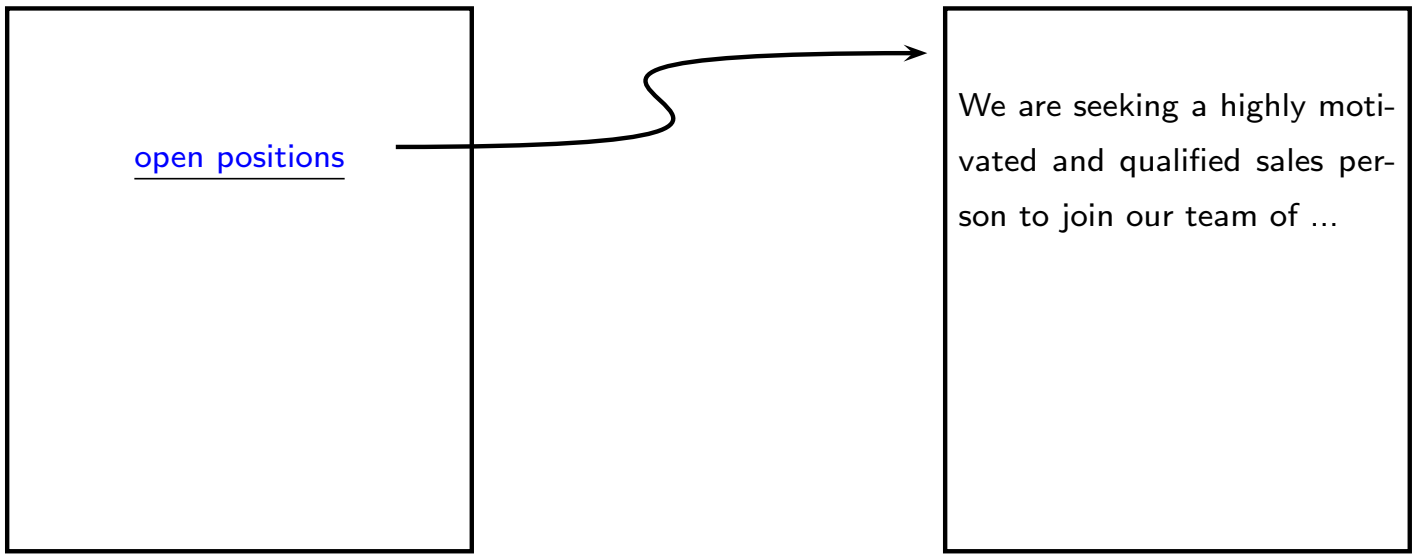
  We have several open positions in our group.

- anchor context:

  For information about open positions click here.

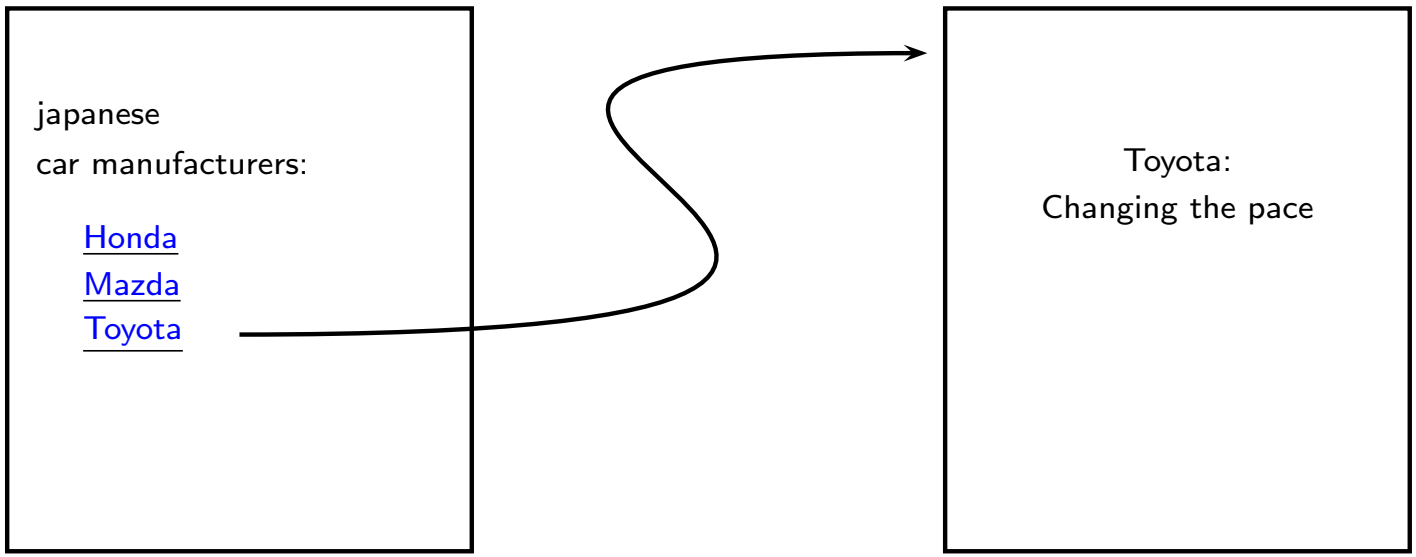
Link Context Analysis

- anchor and target text provide mutual evidence



Link Context Analysis

- sometimes anchor and target are fully complementary



- possible solution: absorbing text from neighboring pages
  - both incoming and outgoing
  - mark the descriptors as belonging to other pages
  - useful for pages with little or no textual information

Link Context Analysis

- combined classifiers can be trained
- problem: possible imbalance between the two information sources
  - one source dominates the overall probability
- co-training: keep the feature spaces disjoint
  - use the scores of one classifier to train the other and vice versa
    $\rightarrow$ semi-supervised training

Link Context Analysis

- supervised classification is infeasible
  - high annotation costs
  - millions of pages
- semi-supervised training: e.g. expectation maximization
  - annotate a small amount of the available data
  - train a classifier
  - classify the remaining data
  - rank the data according to the confidence rating of the classifier
  - select the classification results with highest confidence value
  - add them to the set of already annotated data

Link Structure Analysis

- detection of page clusters (communities)
  - allows thematic disambiguation
  - search space reduction by intersecting cluster
- application of ideas from bibliometry: co-citation
  - two documents are commonly cited by many others
    $\rightarrow$ the documents are somehow related

Link Structure Analysis

- a link can be seen as an implicit recommendation ("citation")
  - assessment of the prestige of a page
    $\rightarrow$ prestigeous pages provide high quality information
- problems
  - link structure is only partially known
    - collecting incoming links by crawling is always incomplete
  - number of citations can be manipulated easily
    - quality of the citing page needs to be considered as well (backlinking)
    - few links from pages with a high rank is better than many links from pages with low rank

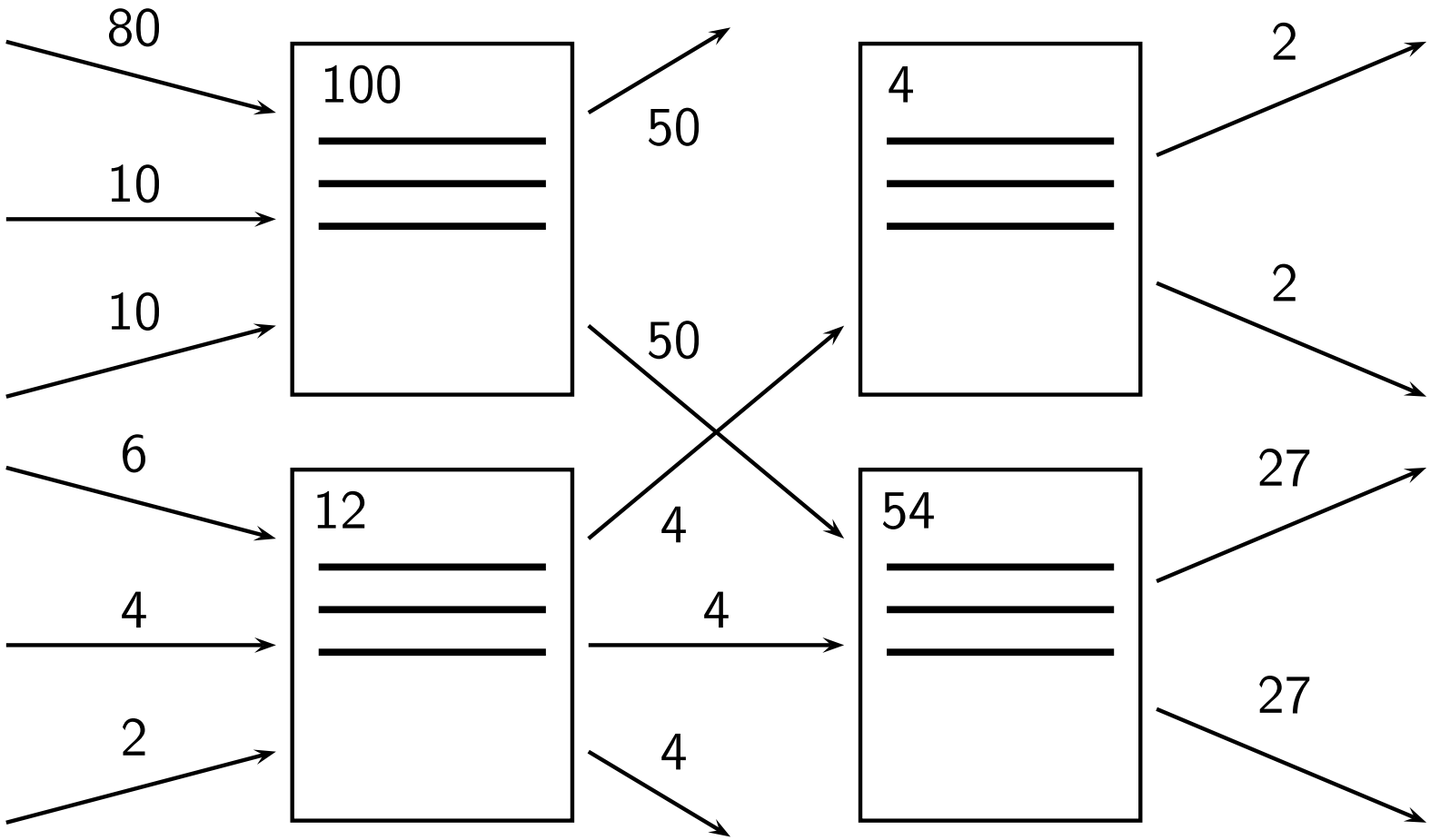
Page Ranking

- PageRank: Page, Brin, Motwani, Winograd (1998)
  $\rightarrow$ Google
- web page: u
- forward links: outgoing from u to other pages $F(u)$
- back links: incoming from other pages to u $B(u)$
- rank: $R(u)$

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{|F(v)|}$$

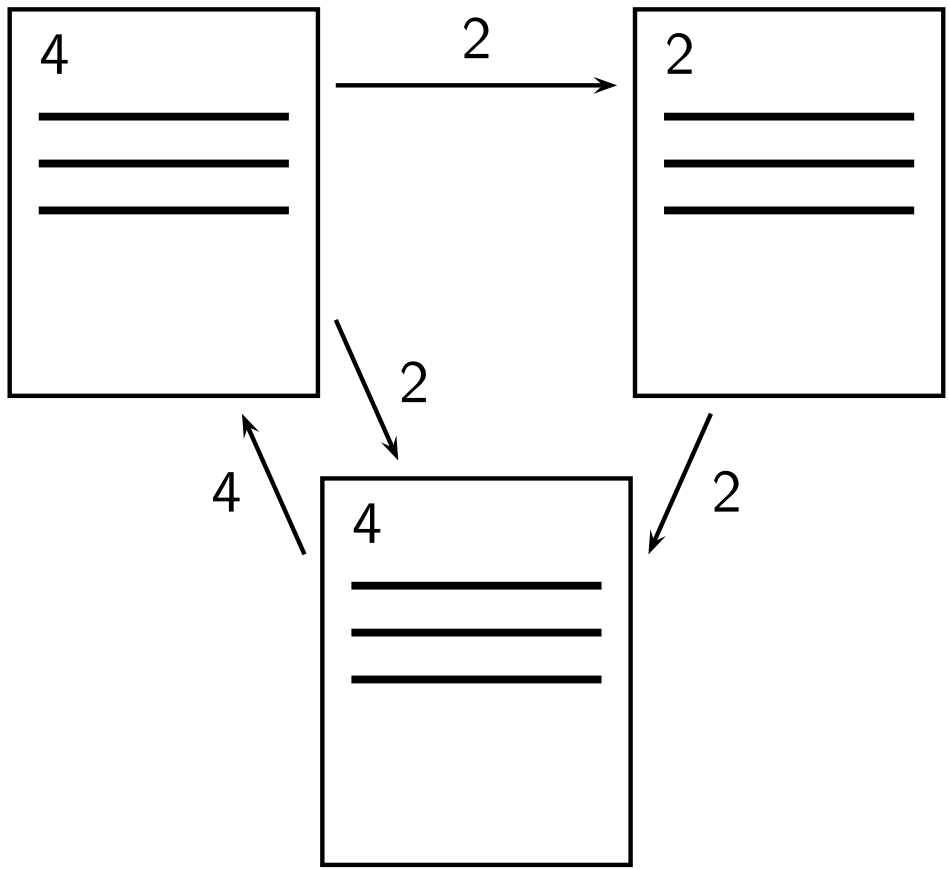
Page Ranking

- pages
  - receive their prestige from incoming links
  - distribute their prestige evenly to the outgoing links



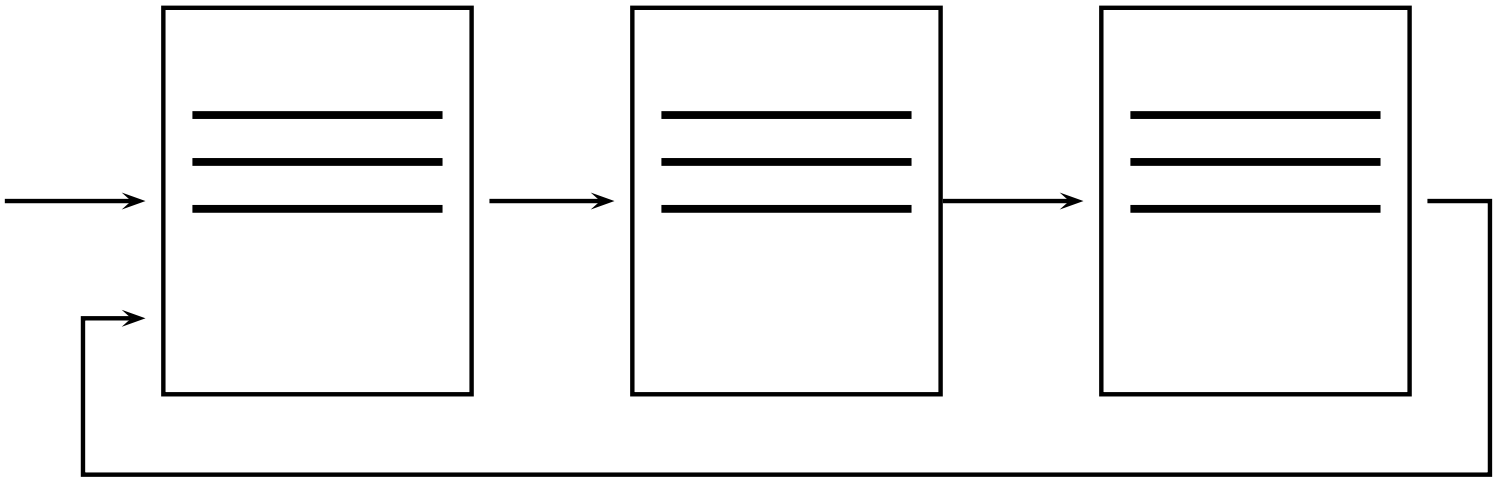
Page Ranking

- recursive algorithm
- repeated approximation until enough stability is achieved
  full convergence is not necessary



Page Ranking

- problem: link cycles without outgoing links (crawler traps)



  $\rightarrow$ unlimited accumulation of weights
- additional decay term included

Page Ranking

- prestige approximated by a random walk model:
  estimation of the probability that a random surfer reaches a particular page
- Google: combination of methods
  - standard text-based distance measures
  - proximity
  - anchor text
  - page ranking

Page Ranking

- adapted page rank

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{|F(v)|} + E(u)$$

- targeted initialisation of $E(u)$
  - personalised rankings: e.g. using personal bookmarks or the links on a users homepage
  - avoids the presentation of (highly linked) address book pages
  - fairly robust against spamming attacks

Page Ranking

- additional applications of PageRank:
  - comparison with access statistics: accessed pages vs. recommended pages
  - backlink prediction: control information for web crawler
  - navigation tools: visualising the importance of links

Page Role Classification

- HITS: hypertext-induced topic selection (Kleinberg, 1998)
- two types of popular web pages:
  - authorities: provide high quality information
  - hubs: link to many authorities
- page role classification based on characteristic structural patterns
  - authority score: sum of hub scores pointing to the page
  - hub score: sum of authority scores to which the page points
- host internal links do not provide evidence of authority (self-citation)
  - need to be ignored

Page Role Classification

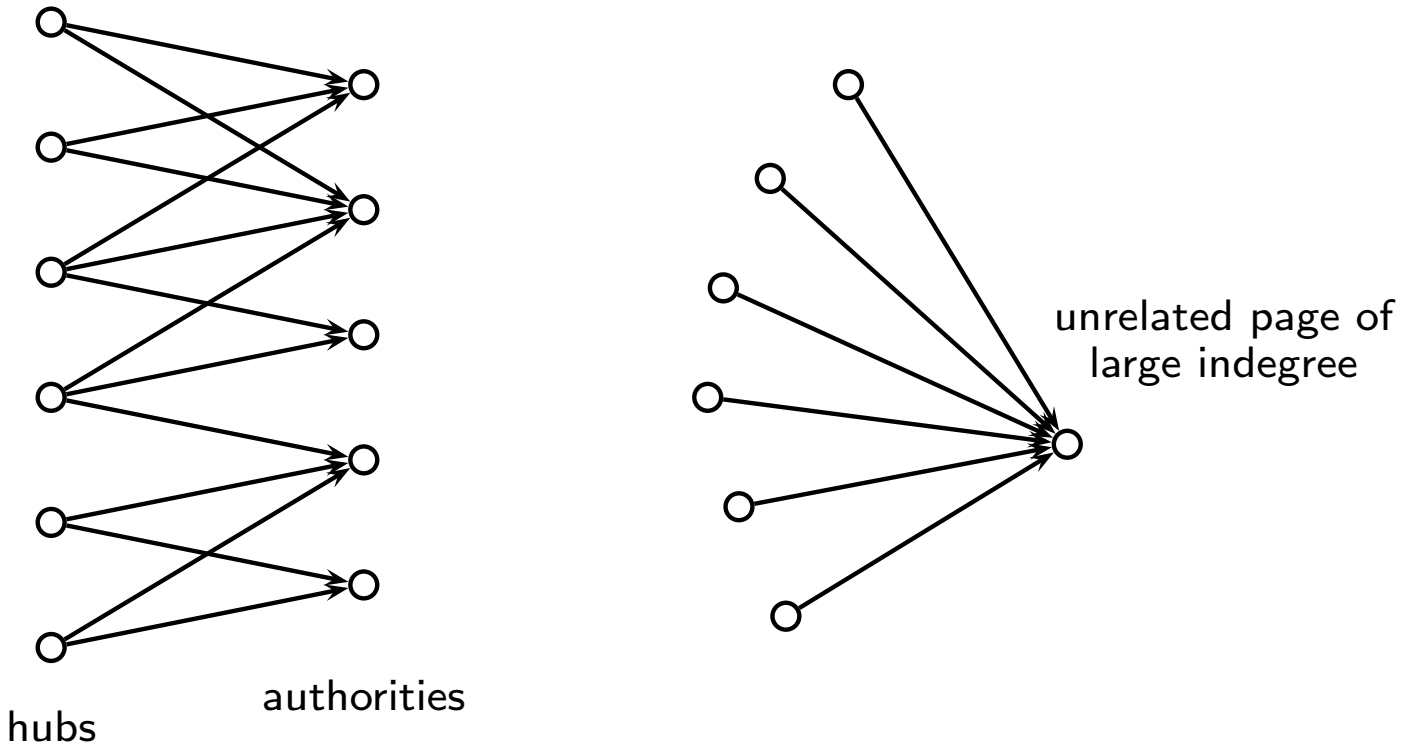
- HITS is topic/query specific
  - prestige values have to be computed online!
- particularly well suited for broad-topic queries
  - environmental protection, smart phones, digital photography, ...
  - usually high number of relevant pages
  - too much to be displayed; filtering required
- hubs are considered more useful, if the user wants to explore a new topic

Page Role Classification

- topic distillation:
  - obtain a set of pages with a text-based search engine (root set)
    - consists of several hundred pages
    - will contain some authorities but not all of them
  - extend the root set by all neighboring pages, both in and out (base set)
    - consists of several thousand pages
    - contains probably most of the authorities
  - eliminate links to the same host

Page Role Classification

- topic distillation (cont.)
  - initialize hub and authority values for each page
  - update hub and authority values iteratively until the network stabilises



  - report top-ranking hubs and authorities

Link Analysis

- other applications of link analysis
  - communication channels and bottlenecks in organizations
  - reputation in social networks
  - opinion leadership in social media
  - trust in e-business
  - detection of fake reviews
  - ...

Link Analysis: Problems

- including the radius-1 neighborhood improves recall
- but leads to topic contamination
- radius-2 neighborhood: most hits are off topic
  - topic generalization: the neighborhood contains pages more general than the original query
    e.g. "movie awards" $\rightarrow$ "movies"
  - topic drift: many sites contain non-semantic links to general pages
    e.g. *This page is best viewed with ...*
- both can be purposefully engineered

Link Analysis: Problems

- spamming link analysis engines
  - setting up fake communities of densely linked web pages
  - linking without semantic connection
  - placing links at public places (discussion forums, community systems, ...)

Database and Information Systems

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

Text Mining

1. Content Extraction
2. Open Domain Question Answering
3. Opinion Mining
4. Trend Detection

Content Extraction

- document retrieval is not enough: a user needs information, not documents
- texts are more than word collections

This site does not contain pornography!

Content Extraction

- full language understanding for unrestricted texts is still infeasible
- new strategy: improvement of results by stepwise accumulation of derived information
  - part-of-speech tagging
  - reference resolution
  - complex attributes
  - language-induced relations
- all information is annotated by an appropriate markup

Content Extraction

- part-of-speech tagging
  - removes some ambiguity, e.g. *can, Bush*
- categories
  - names (persons, places, companies, products, organisations, ...),
  - temporal expressions (date, time)

```
Mr.  <person> Crain </person>, CEO of <company>
GoForMoney </company>, said <date> yesterday </date>
...
```

Content Extraction

- base technologies:
  - word lists (gazetters)
  - (finite-state) pattern `Mr. ?PersonName`
    `?PersonName, CEO of <company>`
    `CEO of ?CompanyName`
- cyclical accumulation of information

Content Extraction

- reference resolution

```
<coref id="1"> HMX </coref>, one of the most serious
  competitors of
<coref id="2"> GlobalView </coref>, gave rise to
  speculations about
<coref id="3" type=pro ref="1"> its </coref> general
  market strategy yesterday. The
<coref id="4" type=def ref="1"> company </coref> is offering
<coref id="5" type=pro ref="4"> its </coref> latest
<coref id="6"> model </coref>, the
<coref id="7" type=name ref="6"> GV200x </coref>, at an
  unusual low price. Said
<coref id="8" type=bridge ref="2"> CEO </coref>
<coref id="9" type=name ref="8"> John Collins </coref>
  ...
```

Content Extraction

- linguistic expressions for reference:
  - Coreference: pronominal, definite
  - nominal: established by a semantic relationship
    is-a, part-of, element-of, name-of, subset-of,
    attribute-of, ...)
- same objects can have different descriptions
  Mr. Bush, the president of the US, the commander in chief, ...
- reference resolution is prerequisite for content extraction
  `selling(HMX,GV200x)`
  `low_price(GV200x)`

Content Extraction

- usually only pronominal reference considered
- preferential criteria:
  - agreement, (Gender, Number)
  - parallelisms (syntactic function, thematic role)
- distance
- typical performance (nominal antecedents): 78-92% precision at 60-64% recall (Baldwin 1997, Stuckardt 2003)
- problem: ambiguity
  Fortgeschrittene Systeme erkennen die Information in der Form, in der sie generiert wird. Sie integrieren sie in das gespeicherte Wissen.

Content Extraction

- complex attribute structures (slot filling)
- sample tasks:
  - terrorism vs. military conflicts vs. criminal events
    location, target, kind, victims, ...

what:	arson attack
where:	Hamburg
victims:	4 serious casualties
damage:	TEuro 800
background:	unknown

  - job offers: company, site, required qualifications, salary, benefits, ...
  - selling events: seller, buyer, goods, amount, price, date, ...
  - booking confirmations: flight-no, date, time, places, category, .
  - customer inquiries: model, type, problem class, ...

Text Mining

1. Content Extraction
2. Open Domain Question Answering
3. Opinion Mining
4. Trend Detection

Open Domain Question Answering

- basis: large text corpora (several million documents)
- task: find a short piece of text which answers a given question e.g.

  Who was Christoph Kolumbus?
  What is a precession?
  When did the 30 Years War start?
  Who was the first man to climb Mount Everest?
- application of pure IR-techniques yields only low response quality: 25.30% acceptable answers

Open Domain Question Answering

- logical inferences doubles the share of correctly answered questions:

question:	*When John Lennon has been killed?*
text:	*On Dec. 8th 1980 John Lennon was shot dead on a street corner in New York city.*
inference:	`shoot(X)` $\rightarrow$ `cause(die(X))`
	`kill(X)` $\rightarrow$ `cause(die(X))`

Competitions

- message understanding conferences (MUC)
- annual evaluation of solutions for information extraction
  - slot filling for special applications
  - named entity recognition
  - reference resolution
  - open-domain question answering
- standardised tasks allow objective comparison od different techniques

Text Mining

1. Content Extraction
2. Open Domain Question Answering
3. Opinion Mining
4. Trend Detection

Opinion Mining

- Bing Liu: Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.
- huge volume of opinionated data in social media on the web
- opinions about
  - products and services (customer satisfaction)
  - political parties and politicians
  - ...
- opinion surveys needed by
  - organizations
  - individual customers, clients, voters, ...

Opinion Mining

- document level: document sentiment classification
- sentence/clause level: subjectivity classification (subjective vs. objective information)
  - different kinds of subjective information can be mixed in a sentence
    The president did exactly the right thing under these terrible circumstances.
- entity/aspect level: feature-based opinion mining
  - opinion consists of a sentiment (positive/negative) and a target (product, aspect of a product)
    Although the service isn't that great, I still love this restaurant.
    The device is quite sturdy but difficult to handle.

Opinion Mining

- sentiment is mostly expressed by means of lexical items or idioms
  good, wonderful, amazing
  bad, poor, terrible
- the sentiment of an item often depends on the context
  The battery charges quite quickly. The battery drains quite quickly.
  The battery lasts quite long. The whole procedure lasts quite long.
- sentiment words do not always express sentiment, e.g. irrealis
  I thought this movie would be as good as the Grinch, but unfortunately, it wasn't.
  but: *But for adults, this movie could be one of the best of the holiday season.*

Opinion Mining

- an opinion can be represented as a 5-tupel
  - the name of an entity
  - an aspect of the entity (part or feature of an entity, or GENERAL)
  - the sentiment on the aspect of the entity (positive/negative/neutral or numerical value)
  - the opinion holder
  - the time point when the opinion was expressed
- not general enough to cover all relevant cases
  - does not distinguish between parts and features
  - no recursive embedding (e.g. features of a part of the entity: *The ink of this printer is too expensive.*
  - relationships between aspects cannot be described: *The view finder and the lens are too close.*

Opinion Mining

- aspects
  - can be given explicitly: *The picture quality is excellent.*
  - or implicitly: *The camera is expensive.*
- opinions
  - can be absolute (regular): *The design is really great.*
  - or comparative: *The design is much cooler than all I have seen so far.*
- opinions
  - can be explicit: *Shelf-life is rather short.*
  - or implicit: *After just two days it started to smell strange.*

Opinion Mining

- supervised document sentiment classification
- train a classifier in supervised mode
- features to be extracted from the document
  - terms and their frequency
  - part-of-speech information
  - sentiment words and phrases
  - sentiment shifters (intensifiers and negators)
  - syntactic relationships

Opinion Mining

- lexicon-based document sentiment classification
- using manually compiled information sources
  - polarity lexicon
  - intensifiers
  - negators
  - irrealis blocking
  - scope heuristics
- individual cues are combined by means of a heuristic measure for sentences and documents

Opinion Mining

- Evaluation
  - quality is extremely domain dependent
  - Twitter postings vs. product reviews vs. forum discussions
  - technical devices vs. hotels vs. films vs. books
  - typical results:

Performance across review types and on positive and negative reviews. (Taboada et al. 2011)

Subcorpus	Epinions 1			Epinions 2		
	Pos-F	Neg-F	Accuracy	Pos-F	Neg-F	Accuracy
Books	0.69	0.74	0.72	0.69	0.77	0.74
Cars	0.90	0.89	0.90	0.80	0.75	0.78
Computers	0.94	0.94	0.94	0.90	0.89	0.90
Cookware	0.74	0.58	0.68	0.79	0.76	0.78
Hotels	0.76	0.67	0.72	0.80	0.70	0.76
Movies	0.84	0.84	0.84	0.76	0.79	0.78
Music	0.82	0.82	0.82	0.83	0.81	0.82
Phones	0.81	0.78	0.80	0.85	0.83	0.84
Total	0.81	0.79	0.80	0.81	0.79	0.80

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

Multimedia Data

- media objects
  - text
  - audio: music, speech
  - graphics: vector graphics, bitmaps
  - video
  - ...

  $\rightarrow$ blobs (binary large objects)
- combinations of media objects: multimedia documents

Multimedia Data

- access to content is increasingly difficult
- characteristics
  - large volume
  - implicit semantics
  - heterogenous types and encoding schemata
  - complex objects
  - various peripheral I/O devices

Multimedia Data

- data preparation
  - decomposition
  - normalization
  - segmentation
- feature extraction
  - graphics: textures, shapes, (stereotypical objects)
  - video: movement vectors
  - audio: spectral or rhythmic properties, (words)
- distance-based similarity retrieval
  e.g. nearest neighbor with R-trees

Multimedia Data

- metadata approach
- describe content and form of multimedia objects
- content-independent: presentation, recording, storing
- content-related: low semantic level (textures, ...)
  - can be extracted automatically
  - mainly for similarity-based retrieval
- content descriptions: high semantic level
  - keyword-based search is possible
  - difficult to be extracted automatically
  - graphics: few stereotypical objects
  - speech: words
    - recognition accuracy 60 ... 80%
      is this sufficient for retrieval purposes?
  - music: simple melodies

Multimedia Data

	meta data	examples
content descriptions (interpretative)	context descriptions	index terms, ontologies, thesauri
	context-related descriptions	identification, location, time, date
	non-textual (1NF) object descriptions	objects, persons, impressions, activities, title
	textual (non 1NF) object decriptions	annotations, transcriptions, scripts, captions
content-related	features	color distribution, texture, sound dynamics, shape
	segment specifications	interval of a stream, contour of a graphical object
content-independent	presentation related recording conditions	QoS, resolution, layout author, recording device
	storage-related	media type, encoding, storage index

(Schmitt 2002)