Text Mining, Information and Fact Extraction Part 5: Integration in Information Retrieval

Marie-Francine Moens
Department of Computer Science
Katholieke Universiteit Leuven, Belgium
sien.moens@cs.kuleuven.be

Information retrieval

- **Information retrieval** (IR) =
  - representation, storage and organization of information items in databases or repositories and their retrieval according to an information need
- **Information items**:
  - format of text, image, video, audio, ...
    - e.g., news stories, e-mails, web pages, photographs, music, statistical data, biomedical data, ...
- **Information need**:
  - format of text, image, video, audio, ...
    - e.g., search terms, natural language question or statement, photo, melody, ...

Is IR needed? Yes

- Large document repositories (archives):
  - of companies: e.g., technical documentation, news archives
  - of governments: e.g., documentation, regulations, laws
  - of schools, museums: e.g., learning material
  - of scientific information: e.g., biomedical articles
  - on hard disk: e.g., e-mails, files
  - of police and intelligence information: e.g., reports, e-mails, taped conversations
  - accessible via P2P networks on the Internet
  - accessible via the **World Wide Web**
  - ...

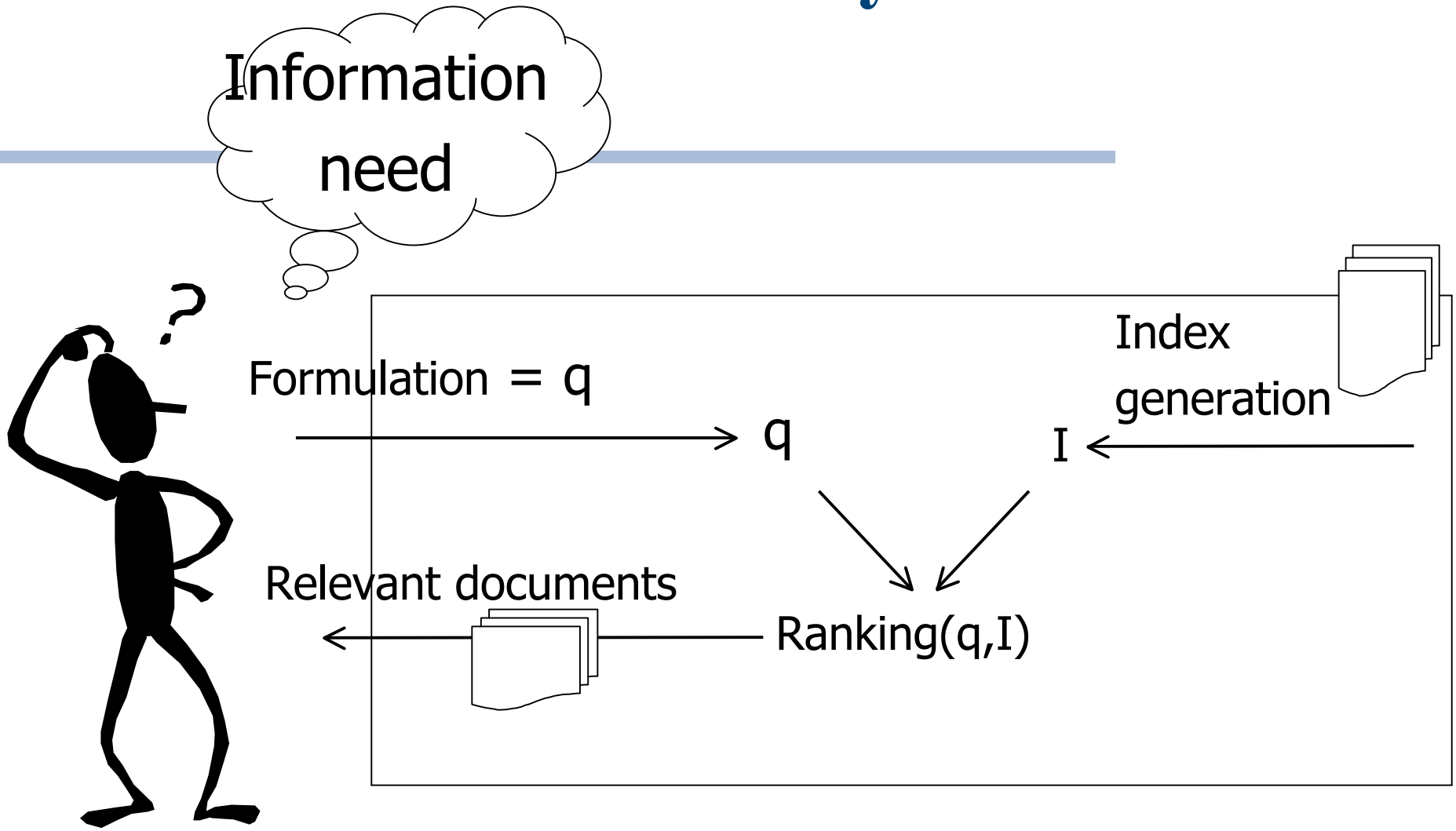
Information retrieval process

- Classical information retrieval system: 3 steps:
  1. generation of a representation of the content of each information item
  2. generation of a representation of the content of the information need of user
  3. the two representations are compared in order to select items that best suit the need

step 1: usually performed before the actual querying

steps 2 and 3 : performed at query time

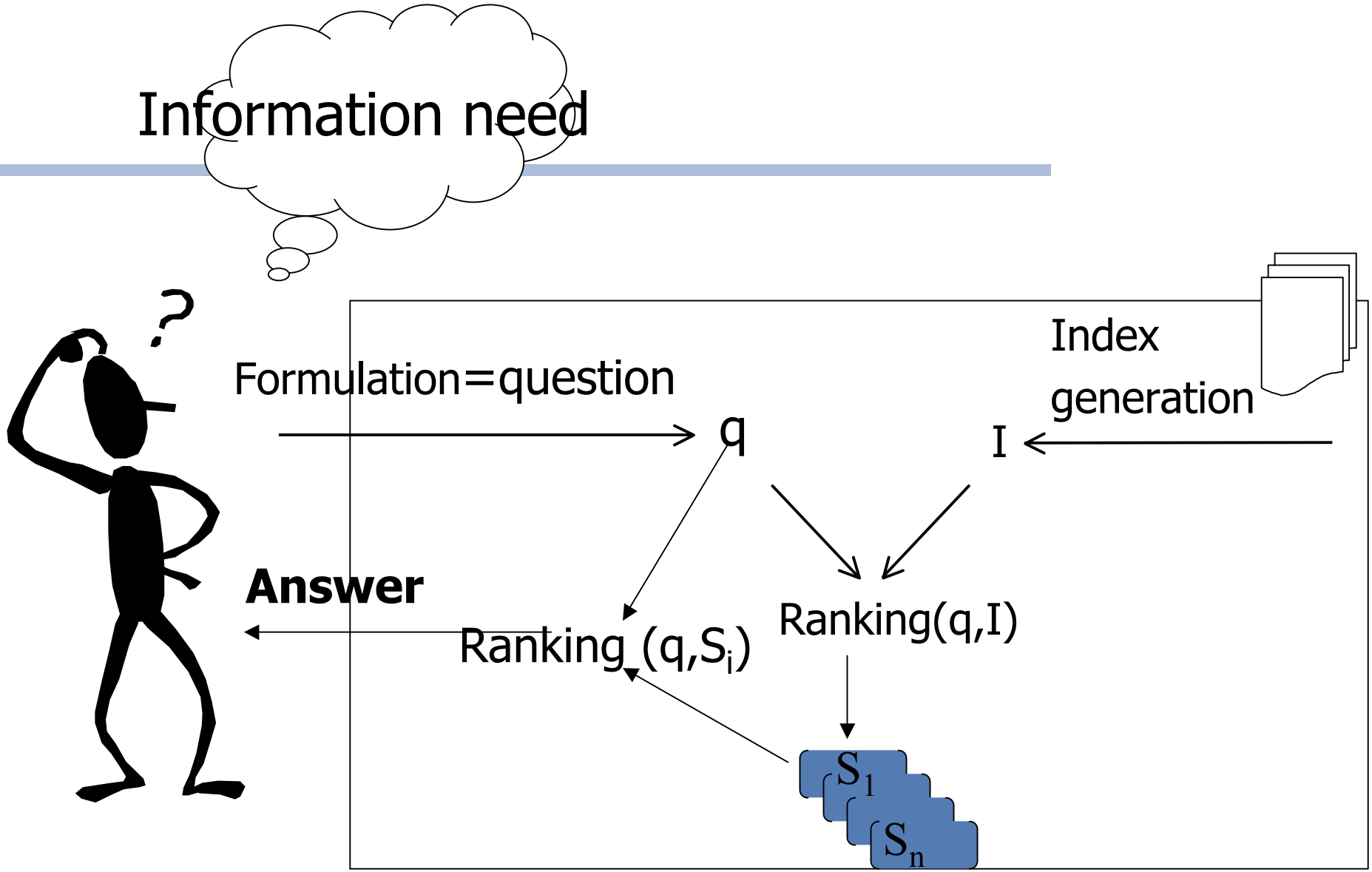
Classical retrieval system



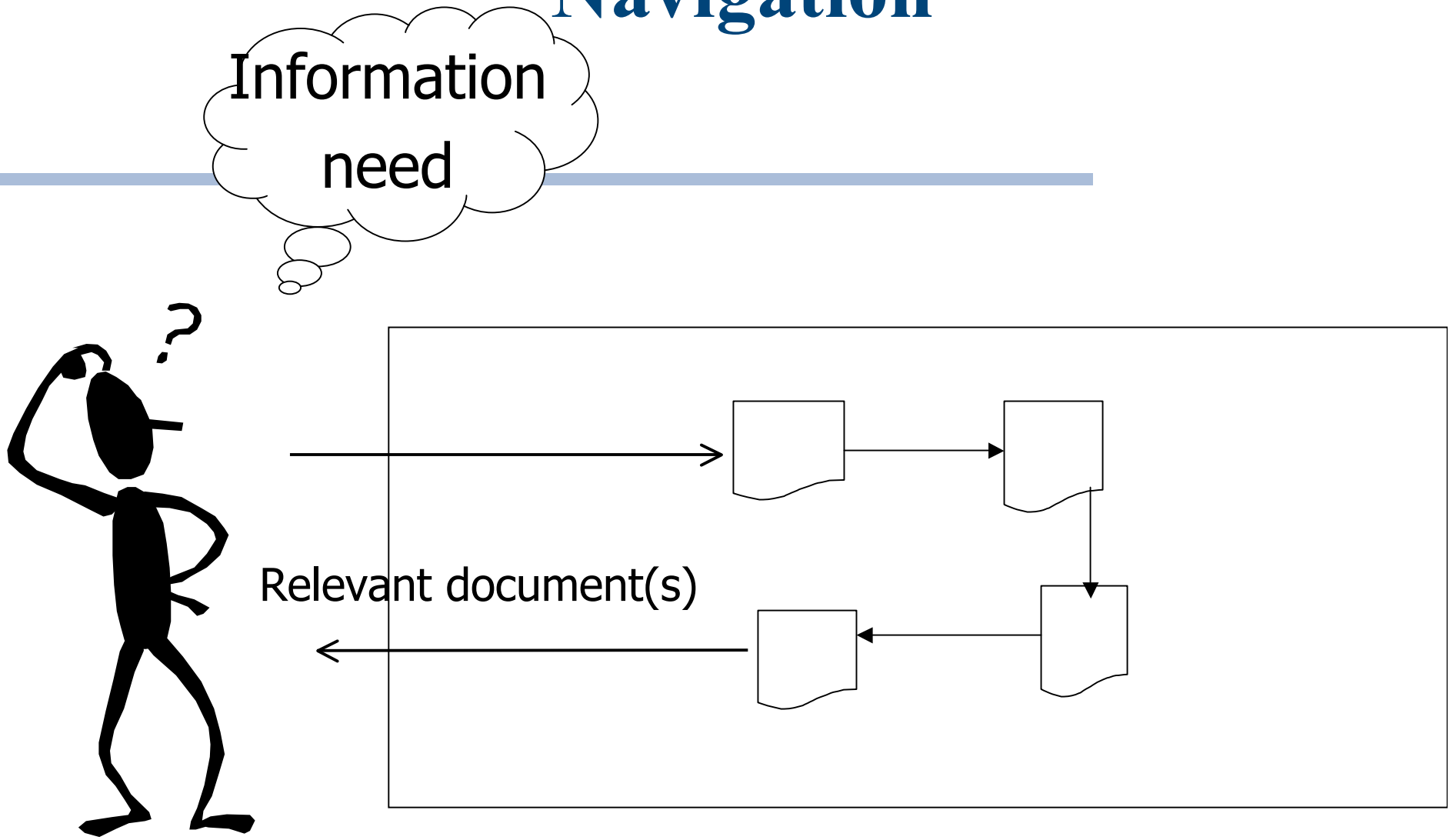
Information retrieval process

- Current retrieval systems
  - information need expressed as:
    - keywords
    - query by example
    - question in natural language
  - Results expressed as:
    - list of documents
    - clusters of documents and visualization of topics
    - short answer to natural language question
- Variant: navigation via linked content
- Future: exploration and synthesis

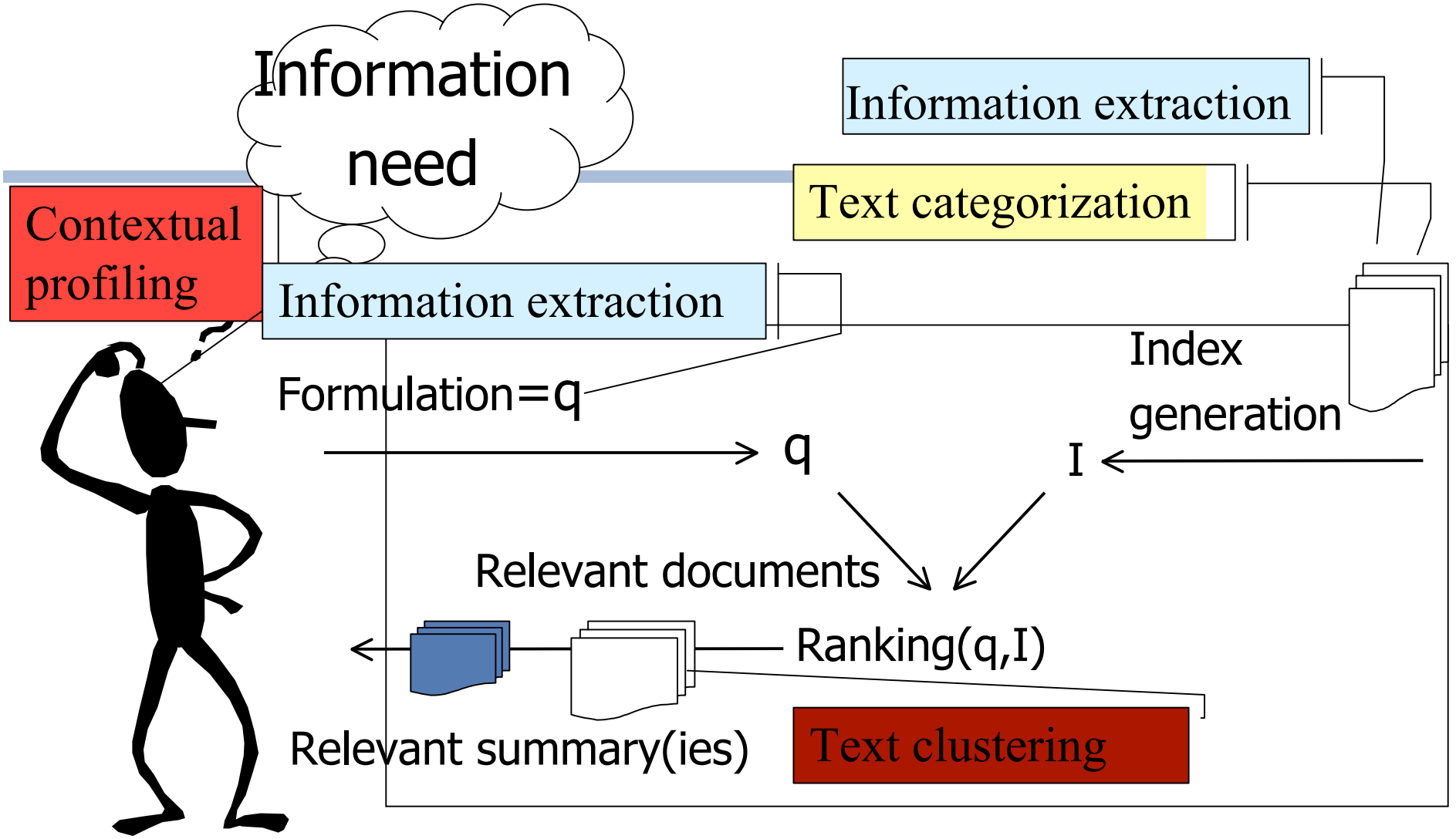
Question answering system



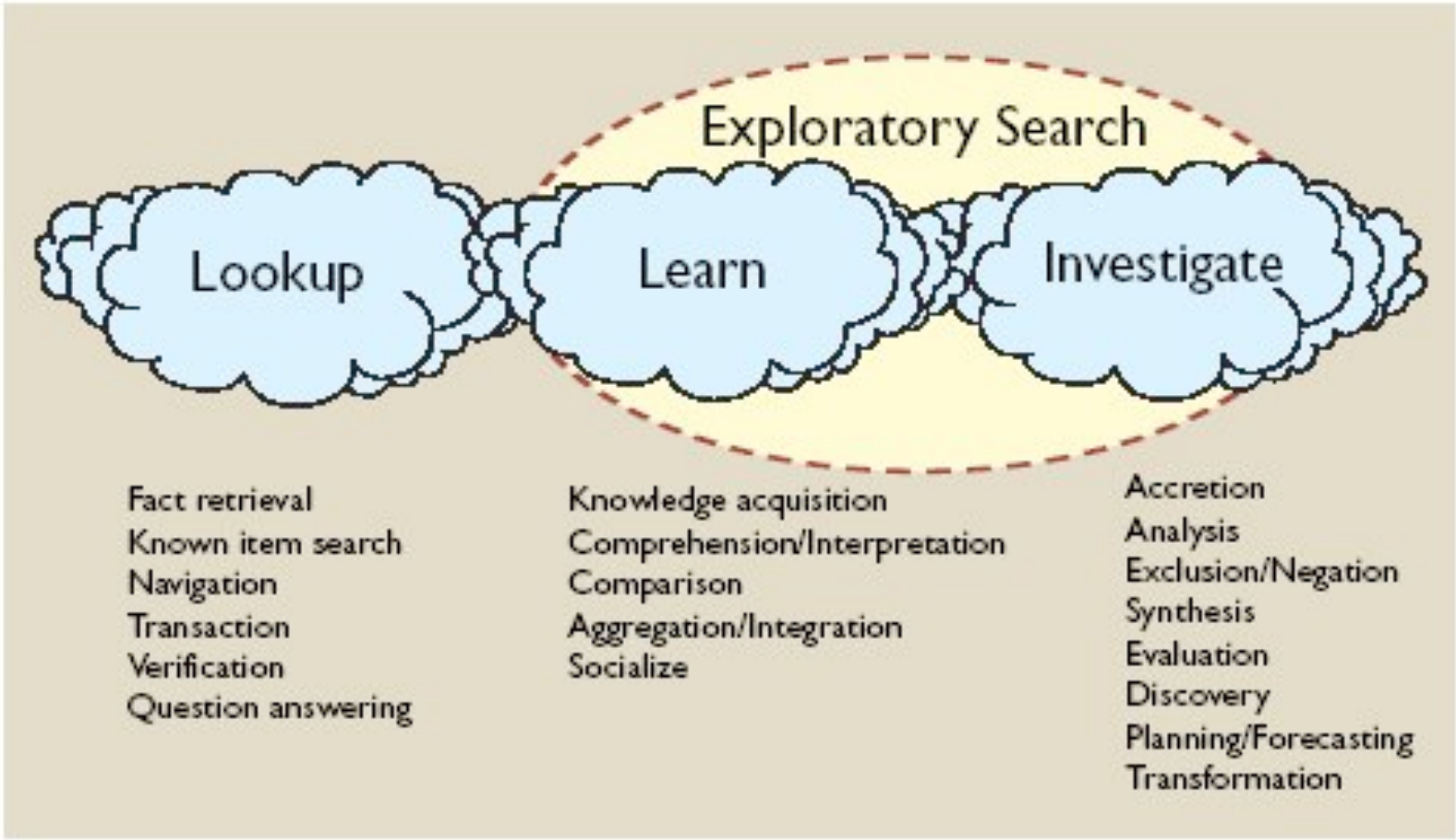
Navigation



Example of assisting tasks



Exploration



[Marchionini ACM 2006]

Why interest in information extraction?

- IR and IE are both established disciplines
- Why this interest now?
  - Catalysts:
    - Question answering
    - Multimedia recognition and retrieval
    - Exploratory search

Overview

- Integration in retrieval models:
  - Language model
  - Entity retrieval
  - Bayesian network
  - Question answering
- Exploratory search
- Interesting research avenues

- **Information retrieval models** (also called ranking or relevance models)
  - defined by :
    - the form used in representing document text and query
    - by the ranking procedure
  - examples are Boolean, vector space, probabilistic models
    - **probabilistic models** incorporate:
      - element of uncertainty

Probabilistic retrieval model

- **Probabilistic retrieval model** views retrieval as a problem of estimating the probability of relevance given a query, document, collection, ...
- Aims at **ranking** the retrieved documents in decreasing order of this probability
- Examples:
  - language model
  - inference network model

Generative relevance models

- Random variables:
  - D = document
  - Q = query
  - R = relevance: $R = r$ (relevant) or $R = \bar{r}$ (not relevant)

- Basic question:
  - estimating:

$$P(R = r \mid D, Q) = 1 - P(R = \bar{r} \mid D, Q)$$

[Robertson & Sparck Jones JASIS 1976] [Lafferty & Zhai 2003]

Generative relevance models

- Generative relevance model: $P(R = r \mid D, Q)$

  is not estimated directly, but is estimated indirectly via Bayes' rule:

$$P(R = r \mid D, Q) = \frac{P(D, Q \mid R = r) P(R = r)}{P(D, Q)}$$

  equivalently, we may use the log-odds to rank documents:

$$\log \frac{P(R = r \mid D, Q)}{P(R = \bar{r} \mid D, Q)} = \log \frac{P(D, Q \mid R = r) P(R = r)}{P(D, Q \mid R = \bar{r}) P(R = \bar{r})}$$

Language model

[Lafferty & Zhai 2003]

- $P(D,Q|R)$ is factored as $P(D,Q|R) = P(D|R)P(Q|D,R)$ by applying the chain rule leading to the following log-odds ratios:

$$\log \frac{P(R=r|Q,D)}{P(R=\bar{r}|Q,D)} = \log \frac{P(Q,D \mid R=r)P(R=r)}{P(Q,D \mid R=\bar{r})P(R=\bar{r})}$$

$$= \log \frac{P(Q \mid D,R=r)P(D|R=r)P(R=r)}{P(Q \mid D,R=\bar{r})P(D|R=\bar{r})P(R=\bar{r})}$$

Bayes' rule and removal of terms for the purpose of ranking

Language model

$$= \log \frac{P(Q \mid D, R = r) P(R = r \mid D)}{P(Q \mid D, R = \bar{r}) P(R = \bar{r} \mid D)}$$

$$= \log \frac{P(Q \mid D, R = r)}{P(Q \mid D, R = \bar{r})} + \log \frac{P(R = r \mid D)}{P(R = \bar{r} \mid D)}$$

The latter term is dependent on D, but independent on Q, thus can be considered for the purpose of ranking.

Assume that conditioned on the event $R = \bar{r}$, the document D is independent of the query Q , i.e.,

$$P(D, Q \mid R = \bar{r}) = P(D \mid R = \bar{r}) P(Q \mid R = \bar{r})$$

Language model

$$\log\frac{P(R=r|Q,D)}{P(R=\bar{r}|Q,D)} = \log\frac{P(Q\mid D,R=r)}{P(Q\mid R=\bar{r})} + \log\frac{P(R=r|D)}{P(R=\bar{r}|D)}$$

$$\overset{rank}{=} \log P(Q|D,R=r) + \log\frac{P(R=r\mid D)}{P(R=\bar{r}\mid D)}$$

Assume that D and R are independent, i.e.,

$$P(D,R) = P(D)P(R)$$

$$\overset{rank}{=} \log P(Q|D,R=r)$$

Language model

- Each query is made of *m* attributes (e.g., n-grams): $Q = (Q_1,\ldots,Q_m)$, typically the query terms, assuming that the attributes are independent given the document and *R*:

$$\log\frac{P(R=r|Q,D)}{P(R=\bar{r}|Q,D)} \overset{rank}{=} \log\prod_{i=1}^{m} P(Q_i|D,R=r) + \log\frac{P(R=r\mid D)}{P(R=\bar{r}\mid D)}$$

$$\overset{rank}{=} \sum_{i=1}^{m} \log P(Q_i|D,R=r) + \log\frac{P(R=r\mid D)}{P(R=\bar{r}\mid D)}$$

- Strictly LM assumes that there is just one document that generates the query and that the user knows (or correctly guesses) something about this document

Language model

- In retrieval there are usually many relevant documents:
  - language model in practical retrieval:

$$P(q_1,\ldots,q_m \mid D) = \prod_{i=1}^{m} P(q_i \mid D)$$

  - takes each document d_j using its individual model $P(q_i|D)$, computes how likely this document generated the request by assuming that the query terms q_i are conditionally independent given the document

  -> **ranking** !

  - needed: smoothing of the probabilities = reevaluating the probabilities: assign some non-zero probability to query terms that do not occur in the document

Language model

- A language retrieval model ranks a document (or information object) *D* according to the probability that the document generates the query (i.e., $P(Q|D)$)
- Suppose the query Q is composed of *m* query terms q_i:

$$P(q_1,\ldots,q_m|D) = \prod_{i=1}^{m}(\lambda P(q_i|D) + (1-\lambda)P(q_i|C))$$

where C = document collection

λ = Jelenik-Mercer smoothing parameter

(other smoothing methods possible: e.g., Dirichlet prior)

Language model

- $P(q_i|D)$ = can be estimated as the term frequency of q_i in d_j upon the sum of term frequencies of each term in D
- $P(q_i|C)$ = can be estimated as the number of documents in which q_i occurs upon the sum of the number of documents in which each term occurs
- Value of λ is obtained from a sample collection:
  - set empirically
  - estimated by the EM (expectation maximization) algorithm
  - often for each query term a λ_i is estimated denoting the importance of each query term, e.g. with the EM algorithm and relevance feedback

Language model

- The EM-algorithm iteratively maximizes the probability of the query given *r* relevant documents $Rd_1,\ldots,Rd_r$:

  init $\lambda_i^{(0)}$ (e.g.: 0.5)

  E-step: $$m_i = \sum_{j=1}^{r} \frac{\lambda_i^{(p)} P(q_i \mid Rd_j)}{(1-\lambda_i^{(p)})P(q_i \mid C) + \lambda_i^{(p)} P(q_i \mid Rd_j)}$$

  M-step: $$\lambda_i^{(p+1)} = \frac{m_i}{r}$$

  Each iteration *p* estimates a new value $\lambda_i^{(p+1)}$ by first computing the E-step and then the M-step until the value $\lambda_i^{(p+1)}$ is not anymore significantly different from $\lambda_i^{(p)}$

Language model

- Allows integrating the translation of a certain content pattern into a conceptual term and the probability of this translation:

$$P(cq_1,\ldots,cq_m|D) = \prod_{i=1}^{m}(\alpha\sum_{l=1}^{k}P(cq_i|w_l)P(w_l|D) + \beta P(cq_i|D) + (1-\alpha-\beta)P(cq_i|C))$$

where cq_i = conceptual terms

w_l = content pattern (e.g., word, image pattern)

- Possibility of building a language model for the query (e.g., based on relevant documents or on concepts of a user's profile)

Language model

- Integration of pLSA or LDA in language model:

$$P(q_1,\ldots,q_m|D) = \prod_{i=1}^{m}(\lambda P(q_i|D) + (1-\lambda)P(q_i|C))$$

where

$$P(q_i|D) = \sum_{k=1}^{K} P(q_i|z_k)P(z_k|D)$$

computed with latent topic model

and K = number of topics (a priori defined)

Language model

- Integrating structural information from XML document
- Computing the relevance of an article *X* nested in a section, which on its turn is nested in a chapter of a statute:

$$P(q_1,q_2,\ldots,q_m|X)=\prod_{i=1}^{m}(\lambda P(q_i|X)+\alpha P(q_i|S)+$$

$$\beta P(q_i|Ch)+yP(q_i|St)+(1-\lambda-\alpha-\beta-\gamma)P(q_i|C))$$

- Allows identifying small retrieval elements that are relevant for the query, while exploiting the context of the retrieval element
- Cf. Cluster based retrieval models: [Liu & Croft SIGIR 2004]

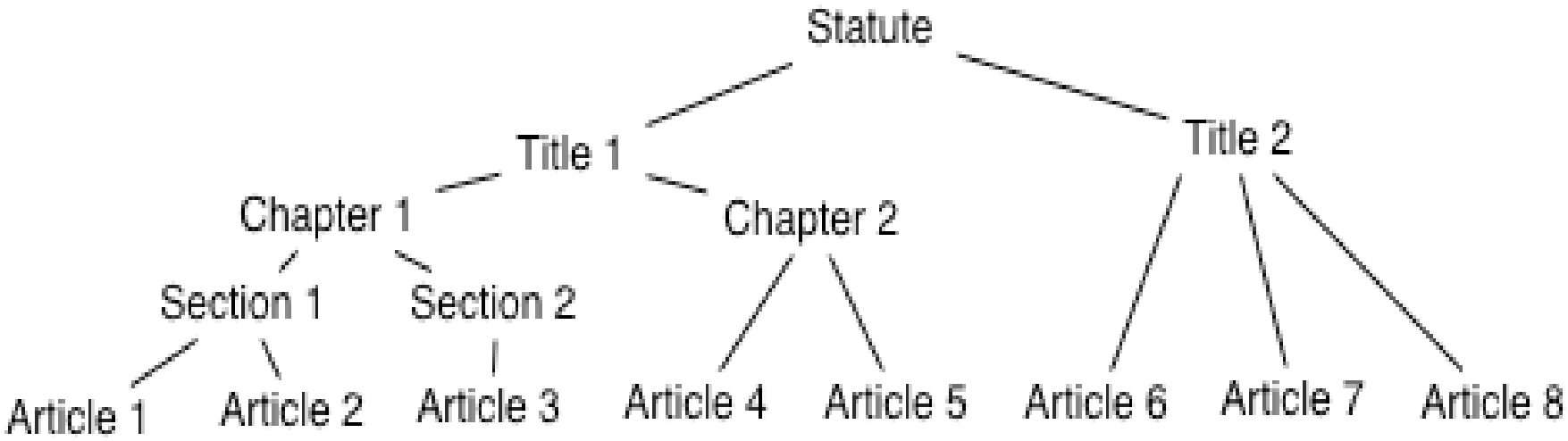


Figure 1. An example structure of a Belgian statute.

Language model

- Advantage:
  - generally better results than the classical probabilistic model
  - incorporation of results of semantic processing
  - incorporation of knowledge from XML-tagged structure (so-called XML-retrieval models)
- Many possibilities for further development

Entity retrieval

- Initiative for the Evaluation of XML retrieval (INEX)
- Entity Ranking track:
  - returning a list of entities that satisfy a topic described in natural language text
- Entity Relation search:
  - returning a list of two entities where each list element satisfies a relation between the two entities: “find tennis player A who won the single title of a grand slam B”

Entity retrieval

- Goal: ranking texts (e_i) that describe entities for an entity search
- **By description ranking:**

$$P(Q|e) = \prod_{t \in Q} P(t|e)$$

$$P(t|e) = (1 - \lambda_c) \frac{tf(t,e)}{|e|} + \lambda_c \frac{\sum_{e'} tf(t,e')}{\sum_{e'} |e|'}$$

where $tf(t,e)$ = term frequency of t in e, $|e|$ is the length of e in number of words, λ_c is a Jelenik-Mercer smoothing parameter

[Tsikika et al. INEX 2007]

Entity retrieval

- **Based on infinite random walk :**
  - Wikipedia texts: links
  - Initialization of $P_0(e)$ and walk: stationary probability of ending up in a certain entity is considered to be proportional to its relevance
  - probability only dependent on its centrality in the walked graph
  - => regular jumps to entity nodes from any node of the entity graph after which the walk restarts:

$$P_t(e) = \lambda_J P(Q|e) + (1 - \lambda_J) \sum_{e' \to e} P(e|e') P_{t-1}(e')$$

    - where λ_J is the probability that at any step the user decides to make a jump and not to follow outgoing links anymore
    - e_i are ranked by $P_\infty(e)$ [Tsikika et al. INEX 2007]

Inference network model

[Turtle & Croft CJ 1992]

- Example of the use of a **Bayesian network** in retrieval
- = directed acyclic graph (DAG)
  - **nodes** = random variables
  - **arcs** = causal relationships between these variables
    - causal relationship is represented by the edge *e* = (*u,v)* directed from each parent (tail) node *u* to the child (head) node *v*
    - parents of a node are judged to be direct causes for it
    - strength of causal influences are expressed by **conditional probabilities**
  - **roots** = nodes without parents
    - might have a **prior probability**: e.g., given based on domain knowledge

Inference network model

- **Document network (DAG)**:
  - contains document representations
  - document (e.g., d_i) represented by:
    - text nodes (e.g., t_j), concept nodes (e.g., r_k), other representation nodes (e.g., representing figures, images)
  - often a document network is once built for the complete document collection:
    - prior probability of a document node

Inference network model

- **Query network (inverted DAG)**:
  - single leaf: information need (Q)
  - information need can have different representations (e.g., q_i) e.g., made up of terms or concepts (e.g., c_l)
  - a query representation can be represented by concepts
- **Retrieval**:
  - the two networks are connected e.g., by their common terms or concepts (attachment) to form the inference or causal network
  - retrieval = a process of combining uncertain evidences from the network and inferring a probability or belief that a document is relevant

Inference network model

- for each document instantiated (e.g. d_j = true (=1), while remaining documents are false(= 0)): the conditional probability for each node in the network is computed
- probability is computed as the propagation of the probabilities from a document node d_j to the query node q

- several evidence combination methods for computing the conditional probability at a node given the parents:
  - e.g., to fit the normal Boolean logic
  - e.g. (weighted) sum: belief a node computed as (weighted) average probability of the parents
- documents are **ranked** according to their probability of relevance

Operators supported by the INQUERY system (University of Massachusetts Amherst, USA) :

#and : AND the terms

#or: OR the terms

#not: negate the term (incoming belief)

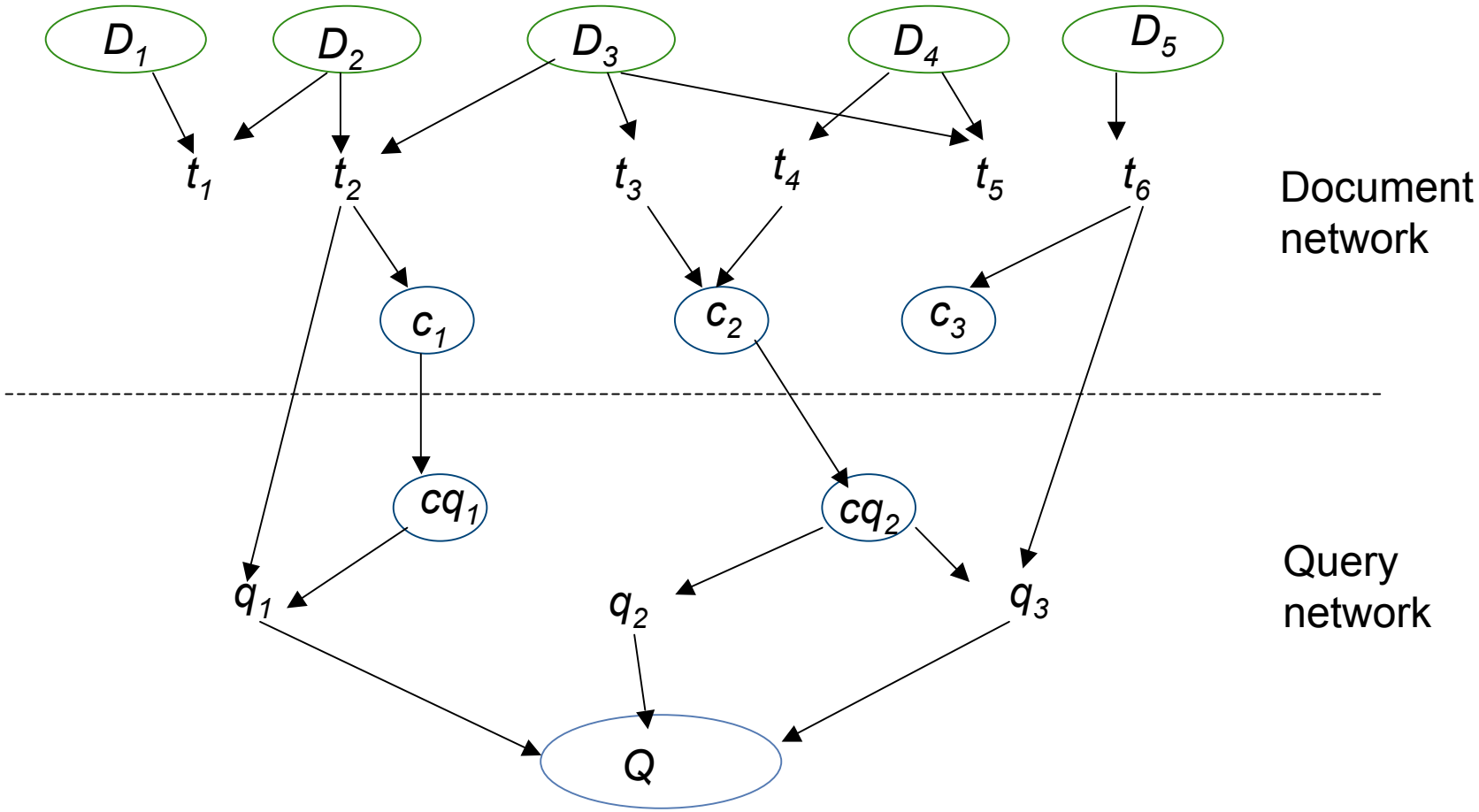
#sum: sum of the incoming beliefs

#wsum: weighted sum of the incoming beliefs

#max: maximum of the incoming beliefs

[Turtle & Croft CJ 1992]

Inference network model



Inference network model

- Advantages:
  - combines multiple sources of evidence and probabilistic dependencies in a very elegant way to suit the general probabilistic paradigm: e.g.,

    P(Query I Document representation, Collection representation, External knowledge,…)
  - in a multimedia database easily integration of text and representations of other media or logical structure
  - easy integration of linguistic data and domain knowledge
  - good retrieval performance with general collections
- **Much new Bayesian network technology yet to be applied !**

Question answering

- **Automatic question answering**:
  - single questions are automatically answered by using a collection of documents as the source of data for the production of the answer
  - interest in the latest Text REtrieval Conferences (TREC)

- Example:

question: “Who is the architect of the Hancock building in Boston?”

answer: “I.M. Pei”

extracted from:

“The John Hancock Tower was completed in 1976 to create additional office space for the John Hancock Life Insurance Co. It was designed by the renowned architect I.M. Pei.”

“Designed by world renowned architect I.M.Pei, the John Hancock Tower is the highest in New England.”

- Example:

Natural language query: “Show me a video fragment where a red car takes a right turn on Saint-John’s square.”

answer:

keyframes of video fragment

extracted from:

video indexed with entities their attributes and relations including spatial and temporal relations

Question answering

- **General procedure:**

1. Analysis of the question
   - selection of key terms for retrieval
   - identification of the question type: e.g. “Who” -> person
   - linguistic analysis of the question: e.g., POS tagging, parsing, recognition of verbs and arguments, semantic role detection and named entity recognition
2. Retrieval of subset of the document collection that is thought to hold the answers and of candidate answer sentences
3. Linguistic analysis of the candidate answer sentences: cf. question

Question answering

- **General procedure:**

1. Analysis of the question
   - selection of key terms for retrieval
   - identification of the **question type**: e.g. “Who” -> person
   - linguistic analysis of the question: e.g., POS tagging, parsing, recognition of verbs and arguments, semantic role detection and named entity recognition
2. Retrieval of subset of the document collection that is thought to hold the answers and of candidate answer sentences
3. Linguistic analysis of the candidate answer sentences: cf. question

Question answering

4. Selection and ranking of answers:
   - candidate sentences are scored usually based on the number of matching concepts and the resolution of an empty slot (**expected answer type**), namely the variable of the question
   - answers can be additionally ranked by frequency
5. Possibly answer formulation

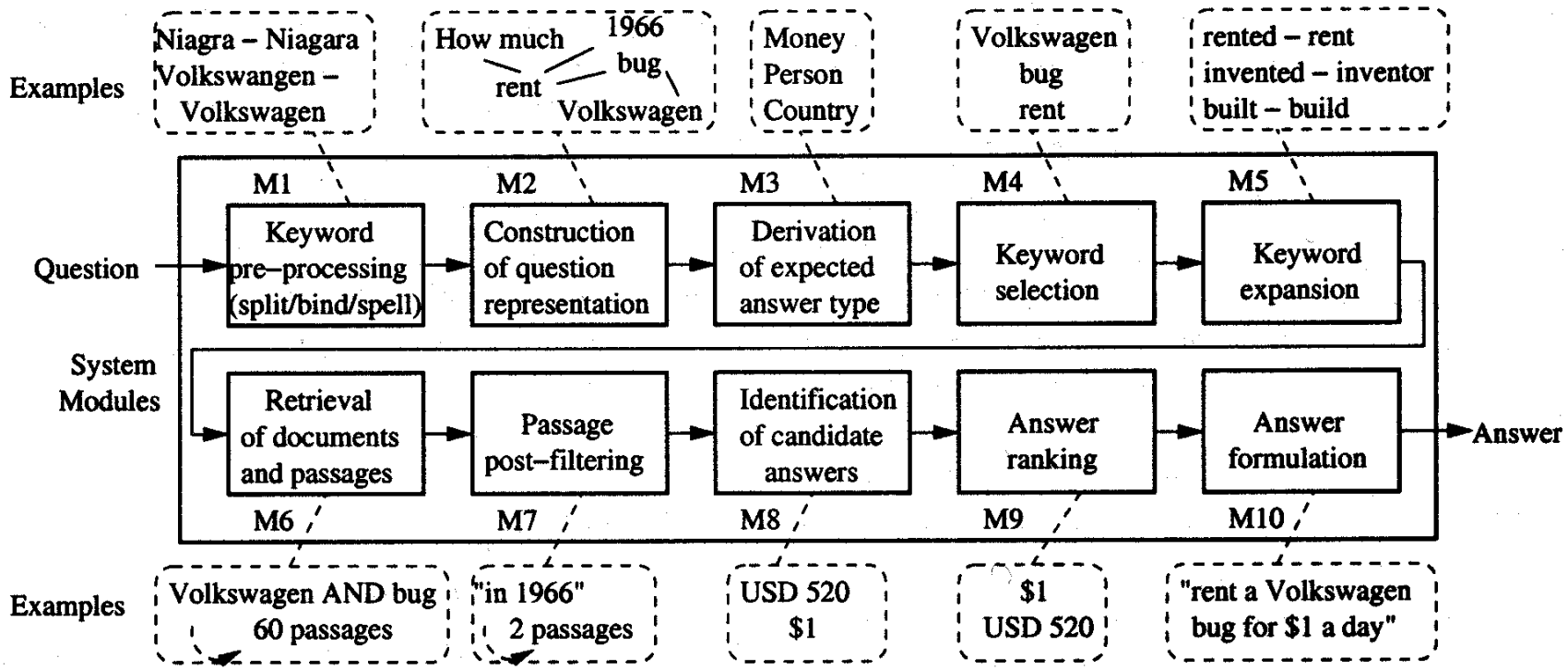


Figure 1: Architecture of baseline serial system (no feedbacks)

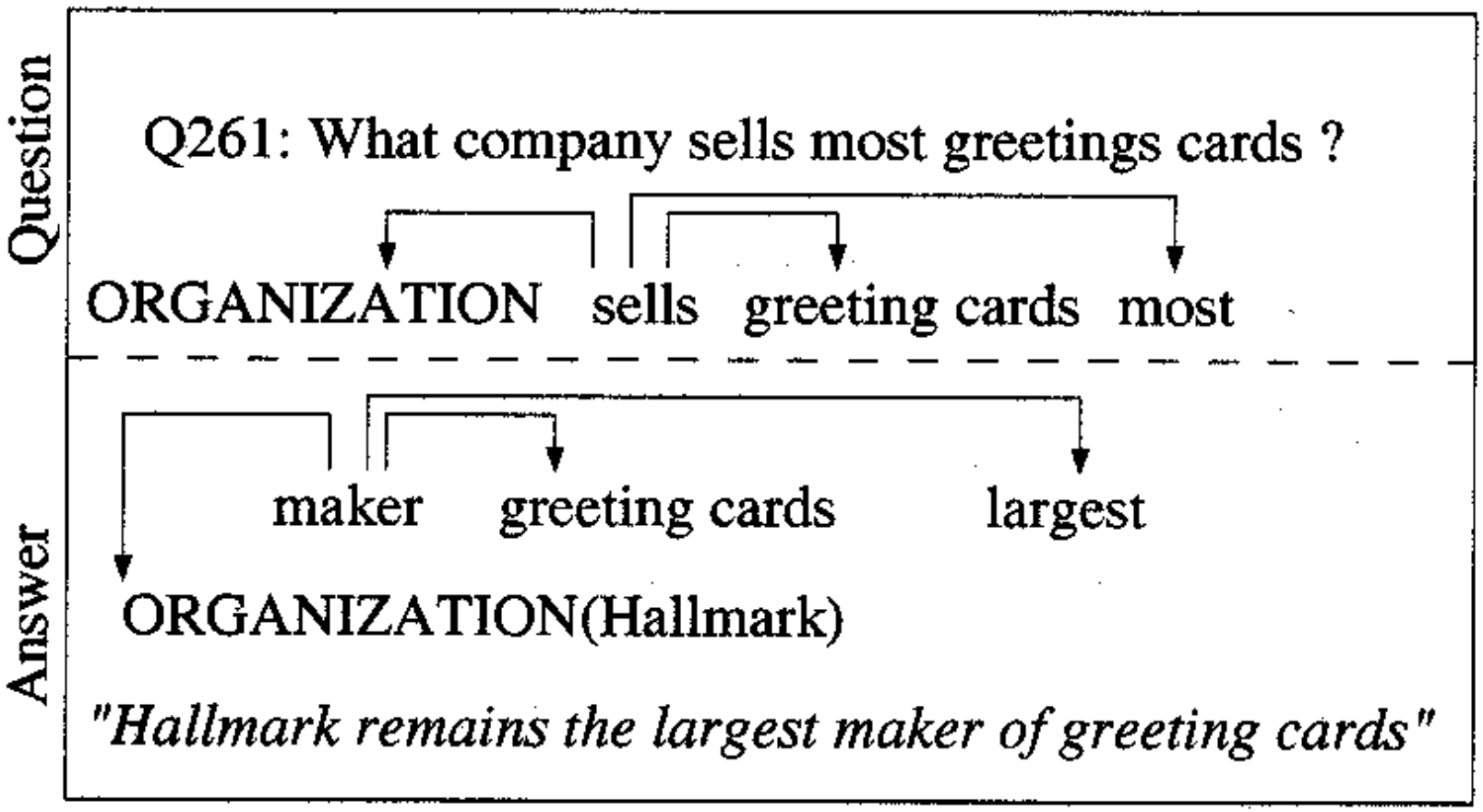
[Moldovan et al. ACL 2002]

Question answering

- **To improve recall** (e.g., no answers found):
  - lexico-semantic alterations of the question based on thesauri and ontologies (e.g., WordNet)
  - morpho-syntactic alterations of the query (e.g., stemming, syntactic paraphrasing based on rules)
  - translation into paraphrases, paraphrase dictionary is learned from comparable corpora
  - incorporation of domain or world knowledge to infer the matching between question and answer sentences

Question answering

- **To improve precision** (e.g., too many answers found):
  - extra constraints: extracted information (e.g., named entity classes, semantic relationships, temporal or spatial roles) from the question and answer sentence must match
  - use of logical representation of question and answer sentences and logic prover selects correct answer



[Pasca & Harabagiu SIGIR 2001]

Question answering

- Difficult task: requires a substantial degree of natural language understanding of the question and of the document texts
- Classes of questions:
  1. **Factual questions**:

     -"When did Mozart die?"

     -answer verbatim in text or as morphological variation

  2. Questions that need **simple reasoning techniques**:

     - "How did Socrates die? ": "die" has to be linked with "drinking poisoned wine"

     -needed: ontological knowledge

Question answering

3. Questions that need **answer fusion from different documents:**

   - e.g., “In what countries occurred an earthquake last year?”

   - needed: reference resolution across multiple texts

4. **Interactive QA systems:**

   - interaction of the user: integration of multiple questions, referent resolution

   - interaction of the system: e.g., “What is the rotation time around the earth of a satellite? “ -> “ Which kind of satellite: GEO, MEO or LEO” ?:

      - needed: ontological knowledge

      - cf. expert system

Question answering

5. Questions that need **analogical reasoning:**
   - speculative questions: “Is the US moving towards a recession?”
   - most probably the answer to such questions is not found in the texts, but an analogical situation and its outcome is found in the text
   - needs extensive knowledge sources, case-based reasoning techniques, temporal, spatial and evidential reasoning
   - very difficult to accomplish due to the lack of knowledge

Question answering results

Table 2: Distribution of errors per system module

Module	Module definition	Errors (%)
(M1)	Keyword pre-processing (split/bind/spell check)	1.9
(M2)	Construction of internal question representation	5.2
(M3)	Derivation of expected answer type	36.4
(M4)	Keyword selection (incorrectly added or excluded)	8.9
(M5)	Keyword expansion desirable but missing	25.7
(M6)	Actual retrieval (limit on passage number or size)	1.6
(M7)	Passage post-filtering (incorrectly discarded)	1.6
(M8)	Identification of candidate answers	8.0
(M9)	Answer ranking	6.3
(M10)	Answer formulation	4.4

[Moldovan et al. ACL 2002]

145	John William King convicted of murder		
	145.1	FACTOID	How many non-white members of the jury were there?
	145.2	FACTOID	Who was the foreman for the jury?
	145.3	FACTOID	Where was the trial held?
	145.4	FACTOID	When was King convicted?
	145.5	FACTOID	Who was the victim of the murder?
	145.6	LIST	What defense and prosecution attorneys participated in the trial?
	145.7	OTHER	
185	Iditarod Race		
	185.1	FACTOID	In what city does the Iditarod start?
	185.2	FACTOID	In what city does the Iditarod end?
	185.3	FACTOID	In what month is it held?
	185.4	FACTOID	Who is the founder of the Iditarod?
	185.5	LIST	Name people who have won the Iditarod.
	185.6	FACTOID	How many miles long is the Iditarod?
	185.7	FACTOID	What is the record time in which the Iditarod was won?
	185.8	LIST	Which companies have sponsored the Iditarod?
	185.9	OTHER	
212	Barry Manilow		
	212.1	FACTOID	What year was he born?
	212.2	FACTOID	How many times has he married?
	212.3	FACTOID	What is the name of the musical that he wrote about the Harmonistas?
	212.4	FACTOID	What music school did he attend?
	212.5	FACTOID	For what female singer was he the musical director and pianist in the 70's?
	212.6	FACTOID	What record label did he sing for in 2000?
	212.7	LIST	List the songs he recorded.
	212.8	OTHER	

Figure 1: Sample question series from the test set. Series 145 has an EVENT as the target, series 185 has a THING as the target, and series 212 has a PERSON as the target.

[Dang et al. TREC 2007]

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lccPA06	Language Computer Corporation (Moldovan)	0.578	0.000	0.000
LCCFerret	Language Computer Corporation (Harabagiu)	0.538	–	0.000
cuhkqaepisto	The Chinese University of Hong Kong	0.390	0.107	0.353
ed06qar1	University of Edinburgh	0.323	0.069	0.294
InsunQA06	Harbin Institute of Technology (HIT)	0.298	0.118	0.353
QACTIS06A	National Security Agency (NSA)	0.266	0.118	0.118
ILQUA1	University of Albany	0.266	0.027	0.059
NUSCHUAQA1	National University of Singapore	0.261	0.000	0.000
asked06c	Tokyo Institute of Technology	0.251	–	0.000
QASCU3	Concordia University (Kosseim)	0.213	0.000	0.000

Table 2: Evaluation scores for runs with the best factoid component.

Run Tag	Submitter	F
lccPA06	Language Computer Corporation (Moldovan)	0.433
cuhkqaepisto	The Chinese University of Hong Kong	0.188
NUSCHUAQA1	National University of Singapore	0.171
FDUQAT15A	Fudan University (Wu)	0.165
QACTIS06C	National Security Agency (NSA)	0.156
LCCFerret	Language Computer Corporation (Harabagiu)	0.148
ILQUA1	University of Albany	0.129
Roma2006run3	University of Rome "La Sapienza"	0.127
csail02	Massachusetts Institute of Technology (MIT)	0.125
InsunQA06	Harbin Institute of Technology (HIT)	0.118

Table 3: Average F-scores for the list question component. Scores are shown for the best run from the top 10 groups.

[Dang et al. TREC 2007]

Run Tag	Submitter	F($\beta = 3$)
ed06qar1	University of Edinburgh	0.250
FDUQAT15A	Fudan University (Wu)	0.223
QASCU3	Concordia University (Kosseim)	0.199
lccPA06	Language Computer Corporation (Moldovan)	0.167
uw574	University of Washington (UW CLMA group)	0.164
Roma2006run3	University of Rome "La Sapienza"	0.164
MITRE2006C	The MITRE Corp.	0.156
QACTIS06C	National Security Agency (NSA)	0.154
NUSCHUAQA3	National University of Singapore	0.150
ISL2	University of Karlsruhe & Carnegie Mellon University	0.150

Table 4: Average F-scores ($\beta = 3$) for the Other questions. Scores are shown for the best run from the top 10 groups.

[Dang et al. TREC 2007]

Question answering in the future

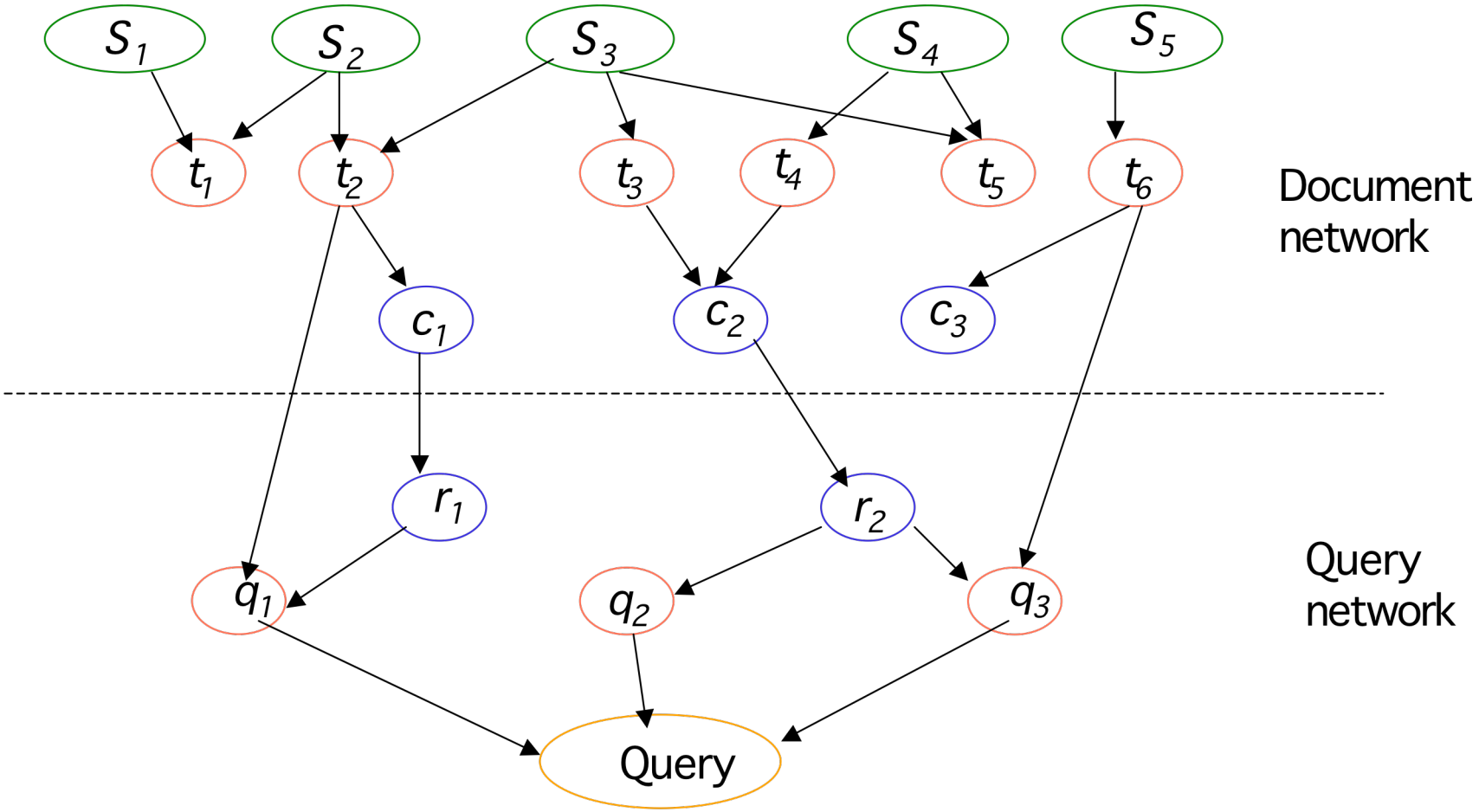
- Increased role of:
  - **information extraction**:
    - Cf. multimedia content recognition (e.g. in computer vision): information also in text will increasingly semantically be labeled
    - Cf. Semantic Web
  - **automated reasoning** (yearly KRAQ conferences):
    - to infer a mapping between question and answer statement
    - for temporal and spatial resolution of sentence roles

Reasoning in information retrieval

- Logic based retrieval models:
  - logical representations (e.g., first order predicate logic)
  - relevance is deduced by applying inference rules
- Inference networks (probabilistic reasoning)
- Possibility to reason across sentences, documents, media, ...:

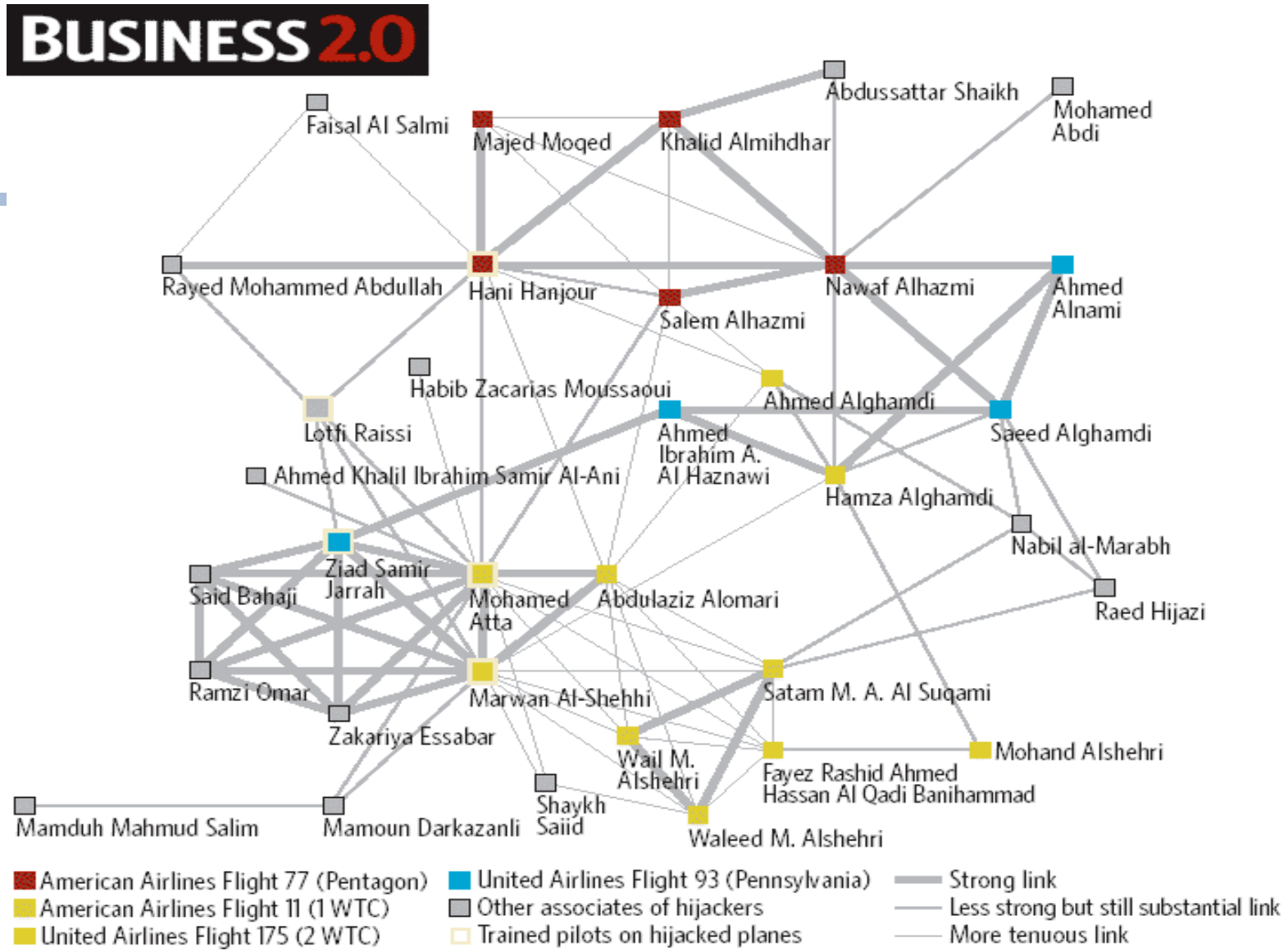
  => **real information fusion**
- **Scalability?**
- ...

An example of an inference network. r_i and c_i represent semantic labels assigned to respectively query and sentence terms. Different combinations of sentences can be activated and their relevance can be computed.



Exploratory search

- Navigation, exploration, analysis of visualized information
- Needed: named entity recognition, relation recognition, noun phrase coreference resolution, ...
- Many applications: intelligence gathering, bioinformatics, business intelligence, ...



Source: COPLINK

We learned

- The early origins of text mining, information and fact extraction and the value of these approaches
- Several machine learning techniques among which context dependent classification models
- Probabilistic topic models
- The many applications in a variety of domains
- The integration in retrieval models

Interesting avenues for research

- Beyond fact extraction
- Extraction of spatio-temporal data and their relationships
- Semi-supervised approaches or other means to reduce annotation
- Linking of content, cross-document, cross-language, cross-media
- Integration in search

AND MANY MORE ...

References

Dang H.T., Kelly, D. & Lin, J. (2007). Overview of the TREC 2007 question answering track. In *Proceedings TREC 2007*. NIST, USA.

Lafferty, J. & Zhai C. X. (2003). Probabilistic relevance models based on document and query generation. In W.B. Croft & J. Lafferty (Eds.), *Language Modeling for Information Retrieval* (pp. 1-10). Boston: Kluwer Academic Publishers.

Liu, X. & Croft, W.B. (2004). Cluster-based retrieval using language models. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49 (4), 41-46.

Moldovan, D. et al. (2002). Performance issues and error analysis in an open-domain question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 33-40). San Francisco: Morgan Kaufmann.

Pasca, M. & Harabagiu, S. (2001). High performance question answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

.

References

Robertson, S. & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129-146.

Tsikrika, T., Serdyukov, P., Rode, H., Westerveld, T., Aly R., Hiemstra, D. & de Vries A.P. (2007). Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In *Proceedings INEX 2007*.

Turtle, H.R. & Croft, W.B. (1992). A comparison of text retrieval models. *The Computer Journal,* 35 (3), 279-290.

Language modeling toolkit for IR: http://www-2.cs.cmu.edu/lemur/