III.6 Advanced Query Types

1. **Query Expansion**
2. **Relevance Feedback**
3. **Novelty & Diversity**

Based on **MRS Chapter 9, BY Chapter 5, [Carbonell and Goldstein ’98] [Agrawal et al ’09]**

1. Query Expansion

- **Query types** in web search according to [Broder '99]
  - **Navigational** (e.g., *facebook*, *saarland university*) [~20%]
    aim to reach a particular web site
  - **Informational** (e.g., *muffin recipes*, *how to knot a tie*) [~50%]
    aim to acquire information present in one or more web pages
  - **Transactional** (e.g., *carpenter saarbrücken*, *nikon df price*) [~30%]
    aim to perform some web-mediated activity
- Problem: Queries are **short** (average: ~2.5 words in web search)
- Idea: **Query expansion** adds carefully selected terms (e.g., from a thesaurus or pseudo-relevant documents) to the query

Thesaurus-Based Query Expansion

- **WordNet** (http://wordnet.princeton.edu) lexical database contains ~200K **concepts** with their **synsets** and conceptual-semantic and lexical **relations**

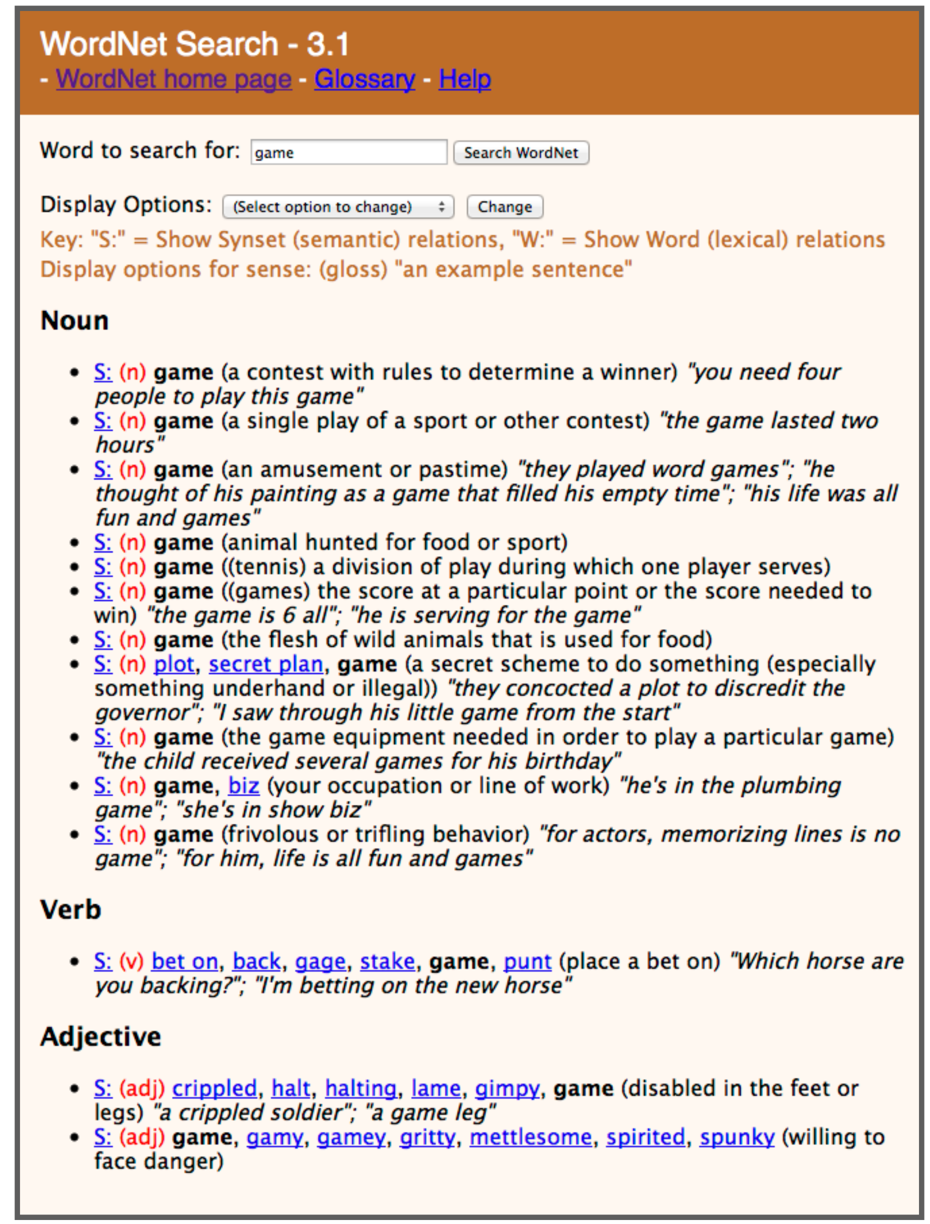
  - **Synonymy** (same meaning)
    e.g.: *embodiment* ⟷ *archetype*

  - **Hyponymy** (more specific concept)
    e.g.: *vehicle* ⟶ *car*

  - **Hypernymy** (more general concept)
    e.g.: *car* ⟶ *vehicle*

  - **Meronymy** (part of something)
    e.g.: *wheel* ⟶ *vehicle*

  - **Antonymy** (opposite meaning)
    e.g.: *hot* ⟷ *cold*



Thesaurus-Based Query Expansion (cont'd)

- **Similarity** *sim*(*u*, *v*) between concepts *u* and *v* based on
  - **co-occurrence statistics** (e.g., from the Web via Google)

$$sim(u, v) = \frac{df(u \wedge v)}{df(u) + df(v) - df(u \wedge v)}$$

  measures **strength of association** (e.g., *car* and *engine*)

  - **context overlap**

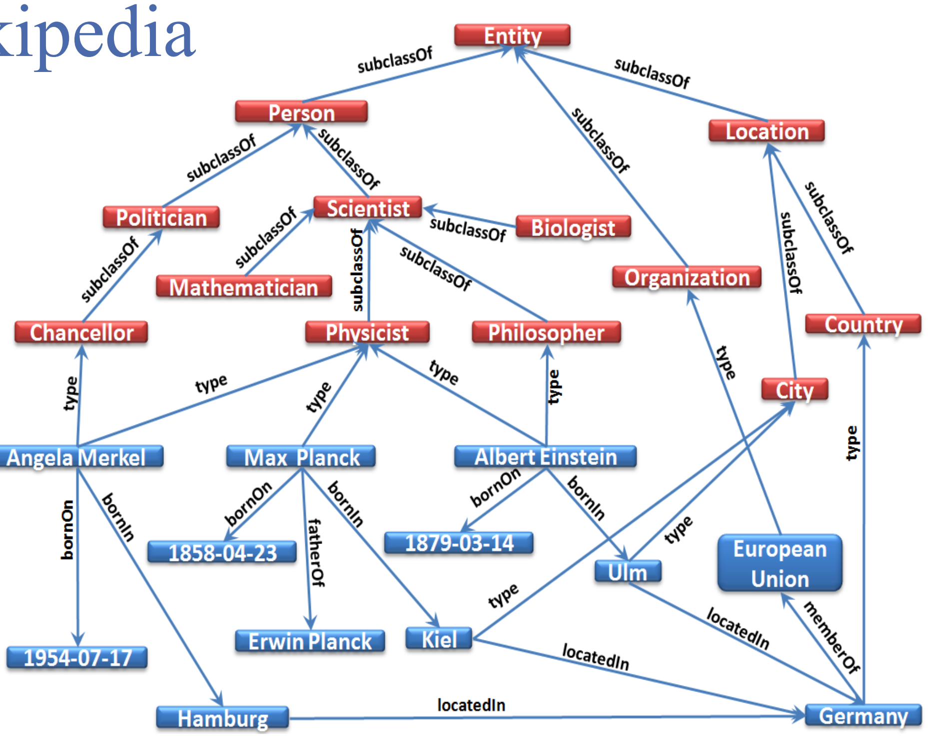
$$sim(u, v) = \frac{|C(u) \cap C(v)|}{|C(u)| + |C(v)| - |C(u) \cap C(v)|}$$

  with *C*(*u*) as the set of terms that **occur often in the context** of concept *u* measures **semantic similarity** (e.g., *car* and *automobile*)

- Expand query by adding **top-*r* most similar terms** from thesaurus

Ontology-Based Query Expansion

- **YAGO** (http://www.yago-knowledge.org) [Hoffart '13]
  - combines knowledge from WordNet and Wikipedia
  - **114 relations** (e.g., marriedTo, wasBornIn)
  - **2.6M entities** (e.g., Albert_Einstein)
  - **365K classes** (e.g., singer, mathematician)
  - **447M facts** (e.g., Ulm locatedIn Germany)

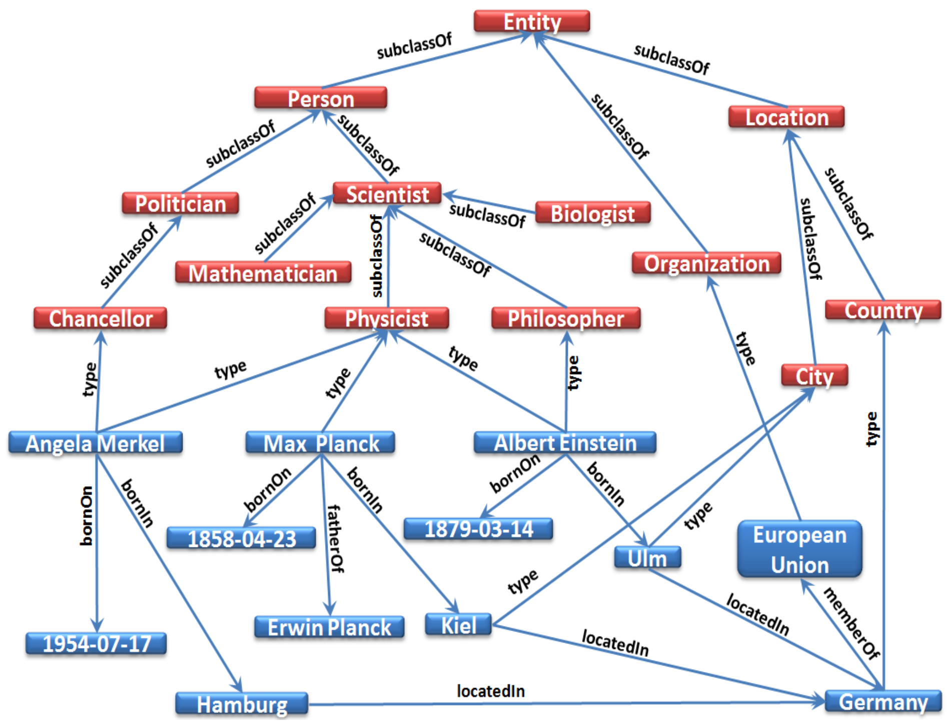


Ontology-Based Query Expansion (cont'd)

- Similarity between classes *u* and *v* based on
  - **Leacock-Chodorow Measure**

$$sim(u, v) = -\log \frac{len(u, v)}{2\,D}$$

with *len*(*u*, *v*) as shortest-path-length between *u* and *v* and *D* as depth of the IS-A hierarchy



  - **Lin Similarity**

$$sim(u, v) = \frac{2\,IC(LCA(u, v))}{IC(u) + IC(v)}$$

with *LCA*(*u*, *v*) as lowest-common-ancestor
and *IC*(*c*) as information content (e.g., number of instances) of class *c*

Local Context Analysis

- Retrieve **top-*n* ranked passages** by breaking initial result documents into smaller passages (e.g., 300 words)
- For each **noun group** *c* (~ concept), compute the similarity *sim*(*q*,*c*) between query *q* and concept *c* using **TF*IDF variant**

$$sim(q,c) = \prod_{t \in q} \left(\lambda + \frac{\log \left(f(c,t)\, idf(c) \right)}{\log\, n} \right)^{idf(t)}$$

$$f(c,t) = \sum_{j=1}^{n} tf(c, p_j) \cdot tf(t, p_j)$$

$$idf(t) = max(1, \frac{log\,(N/np_t)}{5}) \qquad idf(c) = max(1, \frac{log\,(N/np_c)}{5})$$

with constant λ, p_j as the *j*-th passage, and np_t and np_c as the number of passages that contain term *t* and concept *c*, respectively

Local Context Analysis (cont'd)

- Expand query with **top-*m* concepts**. Original query terms receive a weight of 2; the i-th concept added is weighted as $(1 - 0.9 \times i / m)$
- Example: Concepts identified for the query "*What are different techniques to create self induced hypnosis*" include *hypnosis*, *brain wave*, *ms burns*, *hallucination*, *trance*, *circuit*, *suggestion*, *van dyck*, *behavior*, *finding*, *approach*, *study*
- Full details: [Xu and Croft '96]

Global Context Analysis

- Constructs a **similarity thesaurus** between terms based on the intuition that similar terms co-occur in many documents
- TF*IDF variant with **flipped roles** for terms and documents

$$ITF_d = \log\left(\frac{1}{t_d}\right) \qquad \mathbf{t}_d = \frac{(0.5 + 0.5\,\frac{tf_{t,d}}{maxtf_t})\,ITF_d}{\sqrt{\sum_{d'}(0.5 + 0.5\,\frac{tf_{t,d'}}{maxtf_t})^2\,ITF_{d'}^2}}$$

with **inverse term frequency** ITF_d and term vector $\mathbf{t}$

- **Correlation factor** between terms **t** and **t'** is computed as

$$c_{\mathbf{t},\mathbf{t}'} = \mathbf{t} \cdot \mathbf{t}'$$

- Query expanded by **top-*r* terms** most correlated with query terms
- Full details: [Qiu and Frei '93]

2. Relevance Feedback

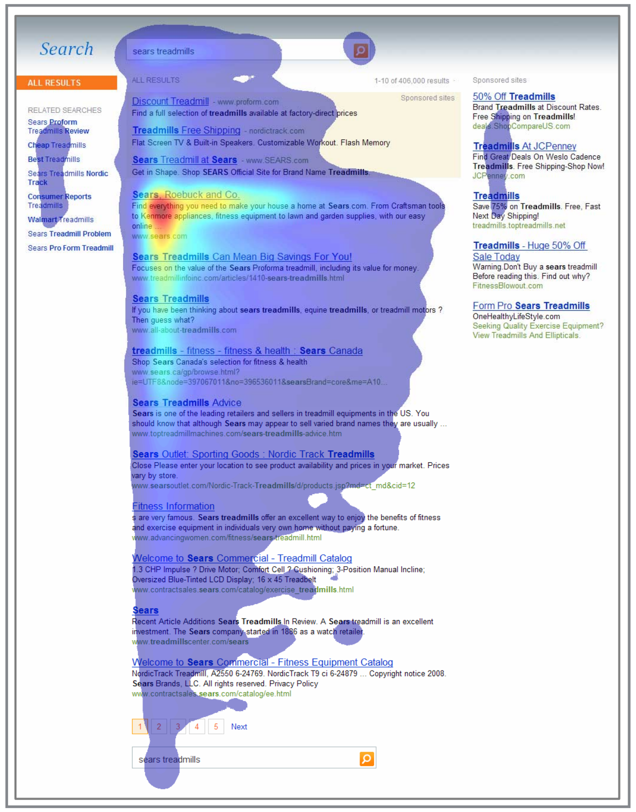
- Idea: Incorporate feedback about **relevant/irrelevant documents**
  - **Explicit relevance feedback** (i.e., user marks documents as +/-)
  - **Implicit relevance feedback** (e.g., based on user's clicks or eye tracking)
  - **Pseudo-relevance feedback** (i.e., consider top-*k* documents as relevant)

- **Relevance feedback** has been considered in all retrieval models
  - **Vector Space Model** (Rocchio's method)
  - Probabilistic IR (cf. III.3)
  - Language Models (cf. III.4)

Implicit Feedback from Eye Tracking

- **Eye tracking** detects area of the screen that is focused by the user in 60-90% of the cases and distinguishes between
  - Pupil fixation
  - Saccades (abrupt stops)
  - Pupil dilation
  - San paths
- **Pupil fixations** mostly user to infer implicit feedback
- **Bias** toward top-ranked search results (receive 60-70% of pupil fixations)
- <u>Possible surrogate</u>: **Pointer movement**

[University of Tampere '07]



[Buscher '10]

Implicit Feedback from Clicks

- Idea: Infer user's preferences based on her clicks in result list

Top-5 Result:

- **Skip-Previous**: $d_2 > d_1$ (i.e., user prefers d_2 oder d_1) and $d_5 > d_4$
- **Skip-Above**: $d_2 > d_1$, $d_5 > d_4$, $d_5 > d_3$, and $d_5 > d_1$

- User study showed **reasonable agreement** with explicit feedback provided for (a) title and snippet of result (b) entire document

- Full details: [Joachims '07]

Rocchio's Method

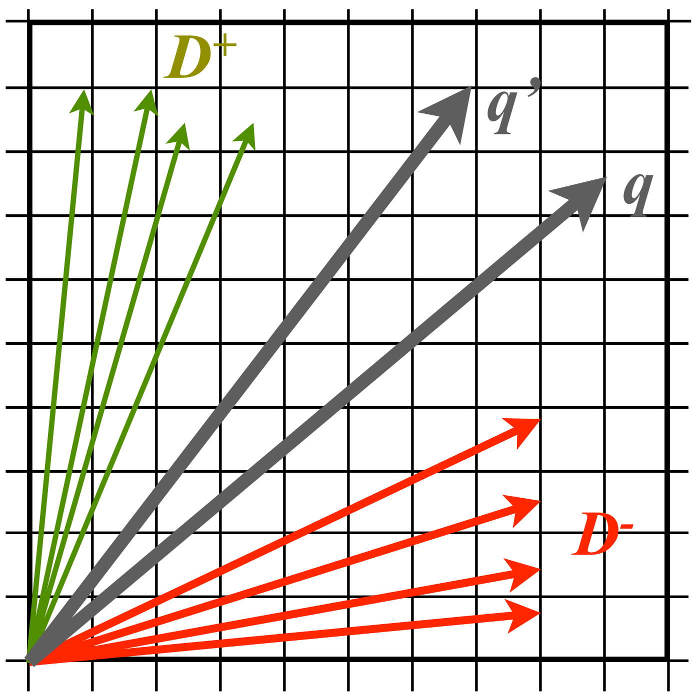
- Rocchio's method considers **relevance feedback** in VSM

- For query $\boldsymbol{q}$ and initial result set D the user provides feedback on **positive documents** $D^+ \subseteq D$ and **negative documents** $D^- \subseteq D$

- **Query vector $\boldsymbol{q}$'** incorporating feedback is obtained as

$$\boldsymbol{q}' = \alpha\, \boldsymbol{q} + \frac{\beta}{|D^+|} \sum_{\boldsymbol{d} \in D^+} \boldsymbol{d} - \frac{\gamma}{|D^-|} \sum_{\boldsymbol{d} \in D^-} \boldsymbol{d}$$

with $\alpha, \beta, \gamma \in [0,1]$ and typically $\alpha > \beta > \gamma$



Rocchio's Method (Example)

	t_1	t_2	t_3	t_4	t_5	t_6	***R***
d_1	1	0	1	1	0	0	**1**
d_2	1	1	0	1	1	0	**1**
d_3	0	0	0	1	1	0	**0**
d_4	0	0	1	0	0	0	**0**

$|D^+| = 2$

$|D^-| = 2$

- Given ***q*** = (1 0 1 0 0 0) we obtain ***q'*** = (0.9 0.2 0.55 0.25 0.05 0) assuming $\alpha = 0.5$, $\beta = 0.4$, $\gamma = 0.3$
- **Multiple feedback iterations** are possible (set ***q*** = ***q'***)

3. Novelty & Diversity

- Retrieval models seen so far (e.g., TF*IDF, LMs) assume that **relevance of documents is independent** from each other
- Problem: **Not a very realistic assumption** in practice due to **(near-)duplicate documents** (e.g., articles about same event)
- Objective: Make sure that the user sees **novel (i.e., non-redundant) information** with every additional result inspected

- Queries are often **ambiguous** (e.g., *jaguar*) with multiple **different information needs behind** them (e.g., car, cat, OS)
- Objective: Make sure that user sees **diverse results** that **cover many of the information needs** possibly behind the query

Maximum Marginal Relevance (MMR)

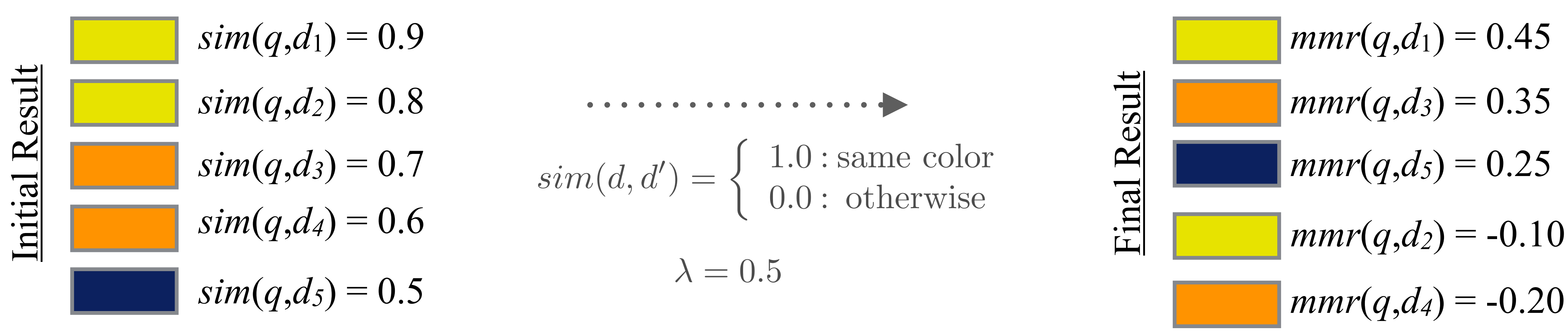
- Intuition: Next result returned d_i should be **relevant to the query** but also **different from the already returned results** $d_1, \ldots, d_{i-1}$

$$\underset{d_i \in D}{arg\,max} \left(\lambda\, sim(q, d_i) - (1 - \lambda) \max_{d_j : 1 \leq j < i} sim(d_i, d_j) \right)$$

with **tunable parameter** λ and **similarity measure** $sim(q,d)$

- Usually implemented as **re-ranking** of top-k query results

- Example:



- Full details: [Carbonell and Goldstein '98]

Intent-Aware Selection (IA-Select)

- Queries and documents are **categorized** (e.g., Technology, Sports)
  - $P(c|q)$ as probability that query q refers to topic c
  - $P(R|d, q, c)$ as probability that document d is relevant for q under topic c
- **IA-Select** determines query result $S \in D$ (s.t. $|S| = k$) as

$$\underset{S}{arg\,max} \sum_{c} P(c|q) \left(1 - \prod_{d \in S} (1 - P(R|d,q,c)) \right)$$

- Intuition: Maximize the probability that **user sees at least one relevant result** for her information need (topic) behind query q
- Problem is ***NP*-hard** but (1-1/e)-approximation, under certain assumptions, can be determined using a greedy algorithm
- Full details: [Agrawal et al. ’09]

Summary of III.6

- **Query expansion**
  counters short query length by adding carefully selected terms based on thesaurus, ontology, global or local context

- **Relevance feedback**
  can be explicit or implicit (e.g., based on clicks or eye tracking) and is applicable in all retrieval models seen so far

- **Novelty & diversity**
  deal with redundancy in query result (e.g., duplicate documents) and ambiguous queries by re-ranking an initial query result

Additional Literature for III.6

- **A. Broder**: *A Taxonomy of Web Search*
  SIGIR 1999

- **R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong**: *Diversifying Search Results*,
  WSDM 2009

- **G. Buscher, S. Dumais, and E. Cutrell:** *The Good, the Bad, and the Random: An Eye-Tracking Study of Ad Quality in Web Search*, SIGIR 2010

- **J. G. Carbonell and J. Goldstein:** *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*, SIGIR 1998

- **J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum**: *YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia*,
  Artificial Intelligence 194:28-61, 2013

- **T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radklinksi, and G. Gay**: *Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search*,
  TOIS 25(2), 2007

- **Y. Qiu and H. P. Frei**: *Concept Based Query Expansion*,
  SIGIR 1993

- **M. Theobald, R. Schenkel, and G. Weikum**: *Efficient and Self-Tuning Incremental Query Expansion for Top-k Query Processing*, SIGIR 2005

- **J. Xu and B. Croft**: *Query Expansion Using Local and Global Document Analysis*,
  SIGIR 1996