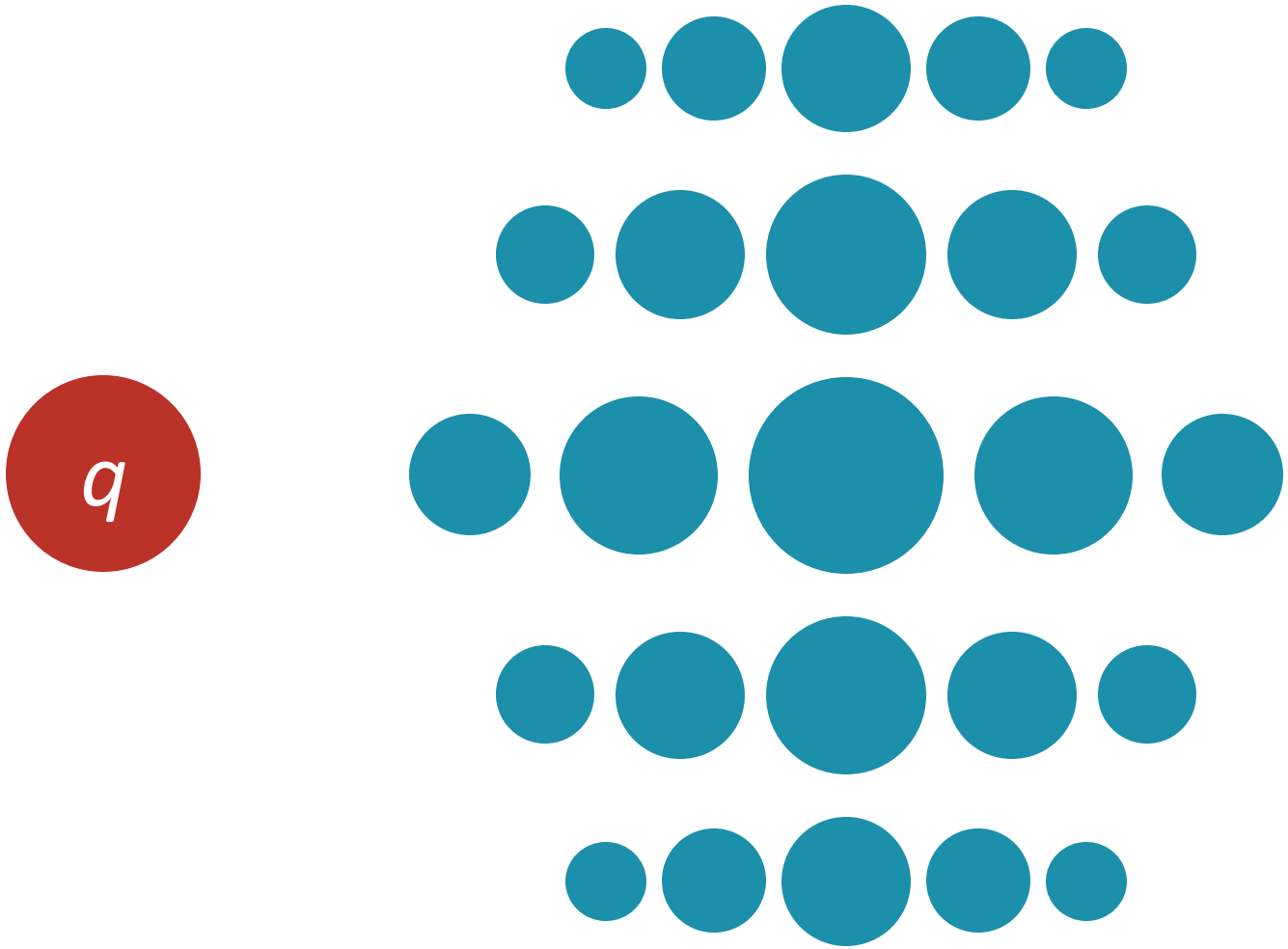
UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Information Retrieval

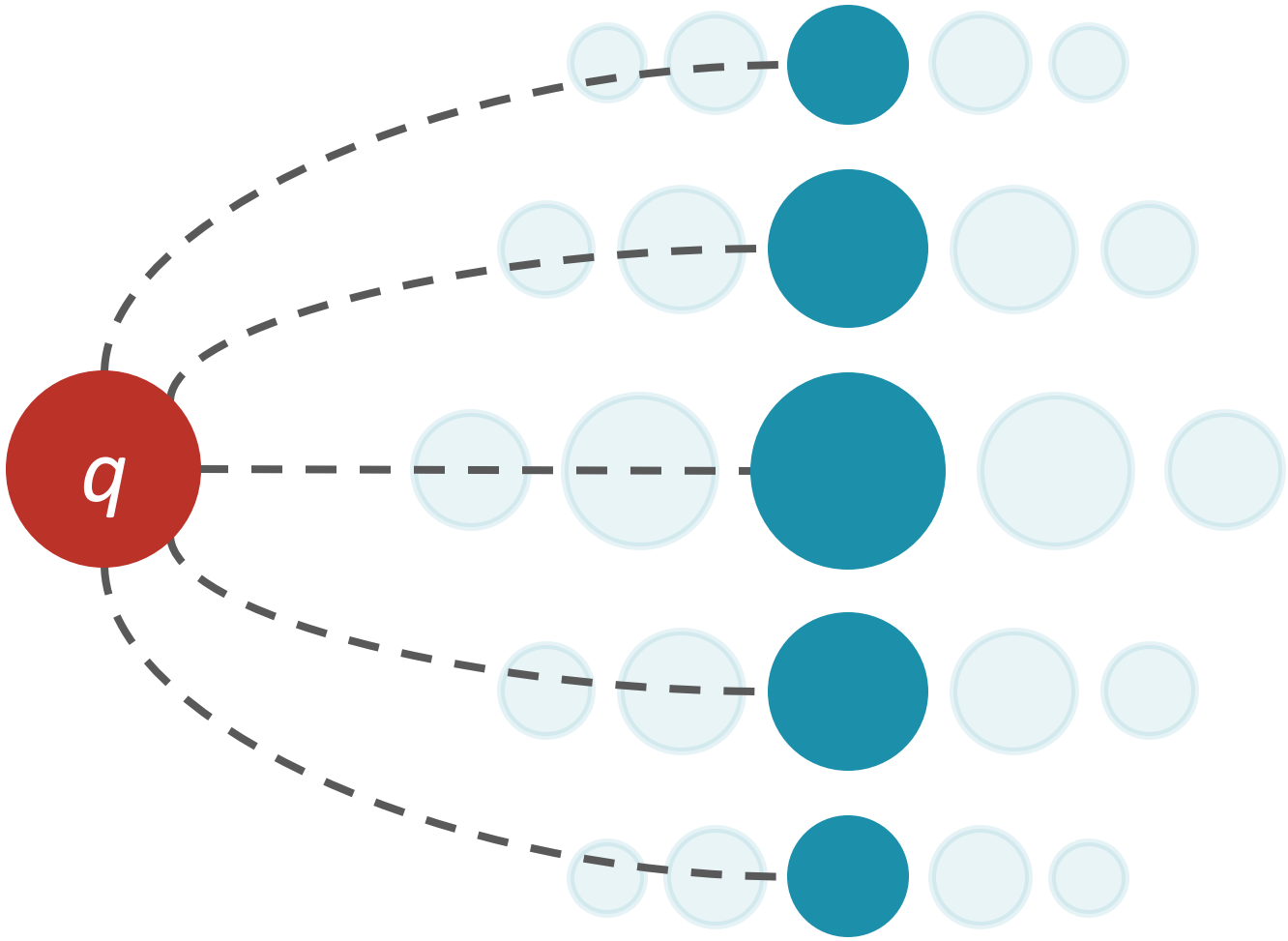
Diversification Models

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

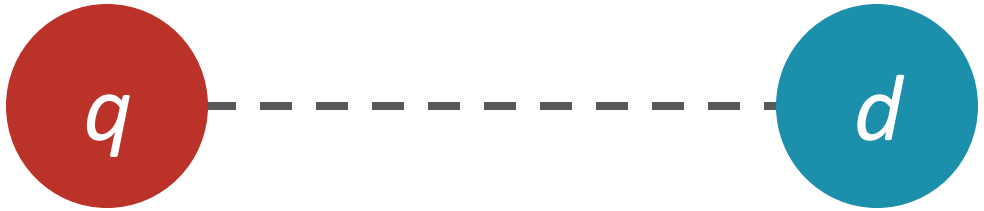
The ranking problem



The ranking problem

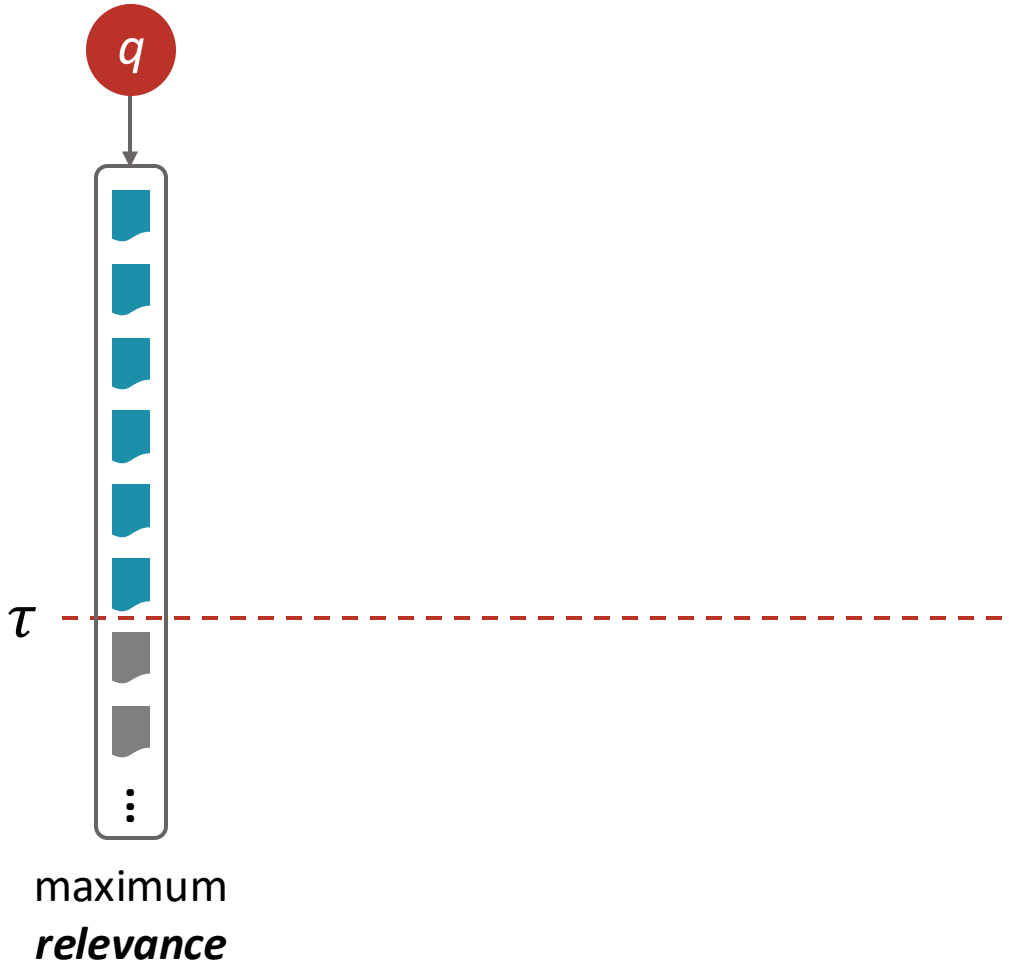


The ranking problem



$$f(q, d)$$

Relevance-oriented ranking



Probability Ranking Principle
(Cooper, 1971; Robertson, 1977)

- Ranking information items by decreasing probability of relevance results in optimal retrieval effectiveness

Ranking optimality

PRP is optimal under certain assumptions
(Gordon & Lenk, 1991, 1992)

- A1. The probability of relevance is estimated with certainty, with no measure of risk
- A2. The probability of relevance is estimated independently for every document

Limiting assumption A1

Assumption

- A1. The probability of relevance is estimated with certainty, with no measure of risk

Limitation: ***ambiguity***

- Information needs and items are ambiguously represented (Turtle & Croft, 1996)

glass

All Images Maps Shopping News More Settings Tools

Sign in

Google

glass

Sign In

All Images Maps Shopping News More Settings Tools

About 2,320,000,000 results (0.72 seconds)

Glass - Wikipedia
https://en.wikipedia.org/wiki/Glass
Glass is a non-crystalline amorphous solid that is often transparent and has widespread practical, technological, and decorative usage in, for example, window ...
Glass ionomer cement: A glass ... · Glass (disambiguation) · History of glass · Sand

material

Glass - X – The Moonshot Factory
https://www.x.company/glass/
Glass Enterprise Edition is a hands-free device, for hands-on workers that removes distractions and helps you focus on what's most important.

smart eyeglasses

Glassdoor Job Search | Find the job that fits your life
https://www.glassdoor.com/index.htm
Search millions of jobs and get the inside scoop on companies with employee reviews, personalized salary tools, and more. Hiring? Post a job for free.
Sign In · Glassdoor Jobs · Companies & Reviews · Know Your Worth

recruiting website

Glass (2019) - IMDb
www.imdb.com/title/tt6823368/
Thriller · The imprisoned Elijah Price holds secrets critical to both David Dunn and Kevin Crumb.

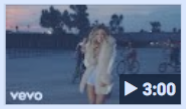
thriller film

Philip Glass
philipglass.com/
Glass holds the Richard and Barbara Debs Composer's Chair at Carnegie Hall for the 2017-2018 season. Highlights will include performances by the Pacific ...

classical composer

Glass | Definition of Glass by Merriam-Webster
https://www.merriam-webster.com/dictionary/glass
Define **glass**: any of various amorphous materials formed from a melt by cooling to rigidity without crystallization: such as — **glass** in a sentence.

word definition

Rachel Platten - Broken Glass - YouTube


https://www.youtube.com/watch?v=b2390GAm4d0
Aug 18, 2017 - Uploaded by RachelPlattenVEVO
Rachel Platten - "Broken **Glass**" (Official Video) Get "Broken **Glass**" when you pre-order her upcoming album ...

pop song

Google Glass 2.0 Is a Startling Second Act | WIRED
https://www.wired.com/story/google-glass-2-is-here/

Glass

Glass is a non-crystalline amorphous solid that is often transparent and has widespread practical, technological, and decorative usage in, for example, window panes, tableware, and optoelectronics. Wikipedia

View 10+ more

Design Furniture

Feedback

Philip Glass (American composer)
Born: January 31, 1937 (age 80), Baltimore, MD
Compositions: Einstein on the Beach, Satyagraha, Akhnat...

Director: M. Night Shyamalan

order her upcoming album ...

Initial release: January 18, 2019
Director: M. Night Shyamalan

Google Glass 2.0 Is a Startling Second Act | WIRED
https://www.wired.com/story/google-glass-2-is-here/ ▾
Jul 18, 2017 - Google **Glass** flopped. Then Alphabet realized that the future of wearables was in factories and warehouses. Welcome to Google **Glass** 2.0.

Glass Enterprise Edition | Glass Explorer Edition | Google Developers
https://developers.google.com/glass/distribute/glass-enterprise ▾
Jul 18, 2017 - **Glass** Partners are authorized to develop and deliver enterprise solutions for **Glass** customers. Learn more here. Except as otherwise noted, the ...

People also ask

Who invented the glass?
What are some of the properties of glass?
How the glass is manufactured?
What are the Google Glasses?

Feedback

Why Google Glass Broke - The New York Times
https://www.nytimes.com/2015/02/05/style/why-google-glass-broke.html
This is the story of Google **Glass**. Before we begin, this is the part in the tale where I should probably explain what Google **Glass** is. Except ...

Boston wants to fight climate change. So why is every new building ...
https://www.bostonglobe.com/ideas/2017/07/14/boston-wants...glass/.../story.html
Yet **glass** buildings also take a lot of energy to heat and cool. When New York started tracking energy use by skyscrapers, the gleaming 7 World ...

Three-dimensional printing of transparent fused silica glass : Nature ...
www.nature.com/articles/nature22061
Glass is one of the most important high-performance materials used for scientific research, in industry and in society, mainly owing to its ...

news + Q&A

Searches related to glass

glass **movie**
what is glass **made of elements**
glass **definition chemistry**
how is glass **made from sand**
glass **chemical formula**
types of glass
glass **imdb**
properties of glass

related searches

Query ambiguity

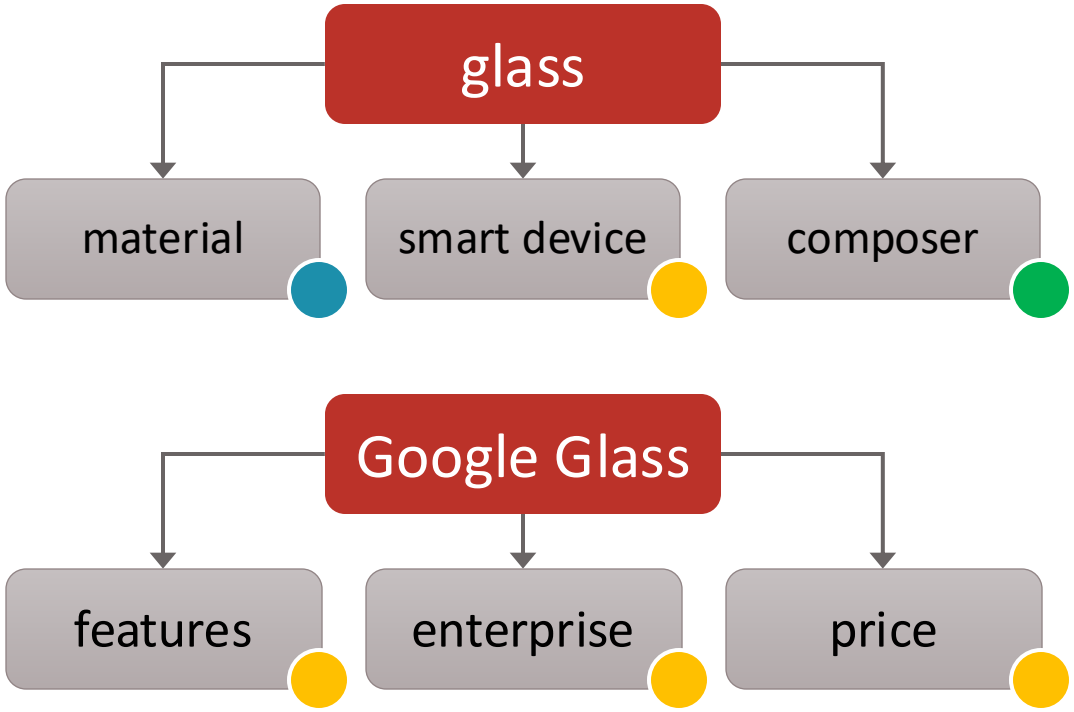
Wikipedia lists over 30 meanings for ‘glass’…

- … but we can only display ‘10 blue links’

Ambiguity is inherent to user queries...

- ... but not all queries are equally ambiguous

Query ambiguity



ambiguous query

↓

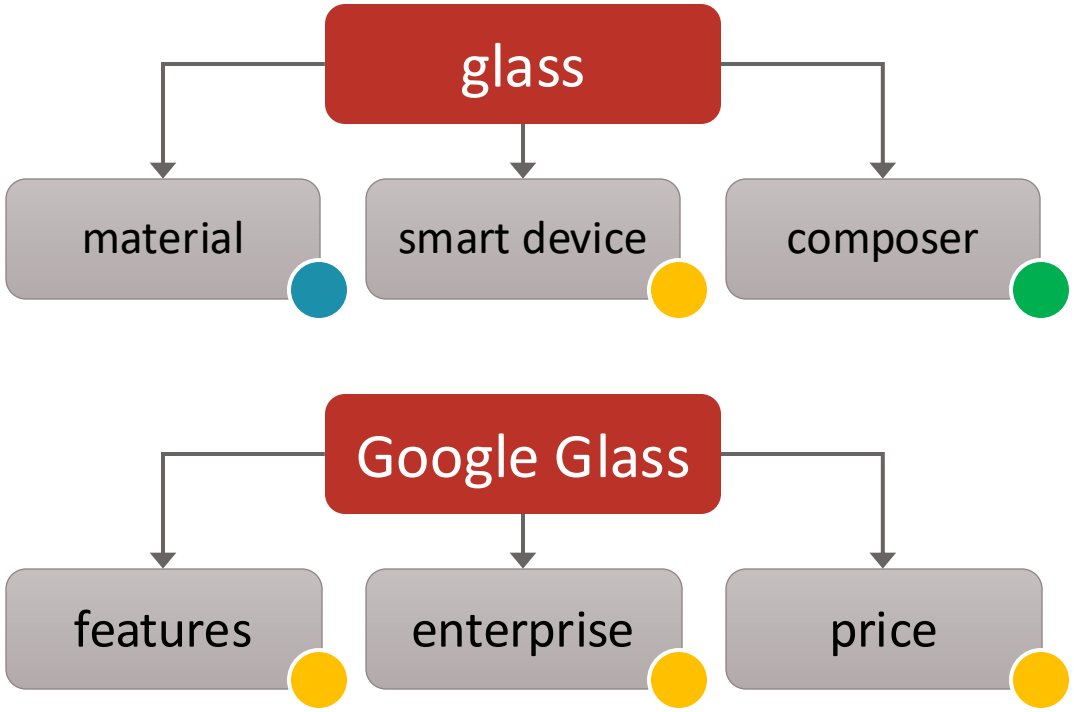
multiple ***interpretations***

underspecified query

↓

multiple ***aspects***

Query ambiguity



ambiguous query

16% of web search queries
(Song, 2009)

underspecified query

every query to some extent
(Cronen-Townsend & Croft, 2002)

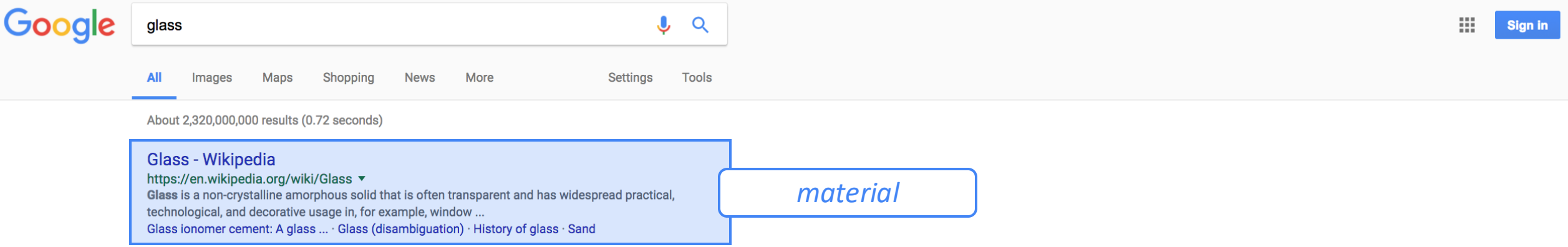
Limiting assumption A2

Assumption

- A2. The probability of relevance is estimated independently for every document

Limitation: ***redundancy***

- "The relationship between a document and a query is necessary but not sufficient to determine relevance" (Goffman, 1964)



Any need for more results about the material?

- Users are unlikely to inspect the results any further once they find something relevant (Craswell et al., 2008)

Diversity and novelty (Clarke et al., 2008)

Diversity

- “the need to resolve ambiguity”
  [in the retrieval request]

Novelty

- “the need to avoid redundancy”
  [in the retrieval response]

Greedy approximation

Diversification is NP-hard (Agrawal, 2009)

- No efficient exact solution

Constant-factor ($1 - 1/e \approx 0.632$) approximation

- Iteratively select a document that covers the most aspects yet uncovered by the previous documents

Greedy approximation

$D \leftarrow \emptyset$

while $|D| < \tau$ **do**

$d^* \leftarrow \text{argmax}_{d \in R} f(q, d, D)$

$R \leftarrow R \setminus \{d^*\}$

$D \leftarrow D \cup \{d^*\}$

end while

return D

Greedy approximation

Approximation effective in practice (Carterette, 2009)

- Minor deviations from the optimal solution

Most approaches focus on producing effective diversification objectives $f(q, d, D)$

A departure from the PRP

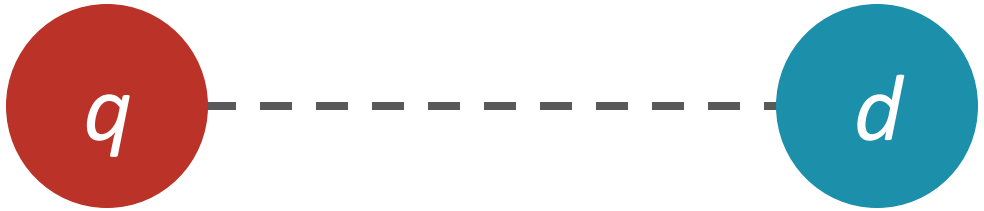
Relevance no longer certain

- Query ambiguously convey multiple needs

Relevance no longer independent

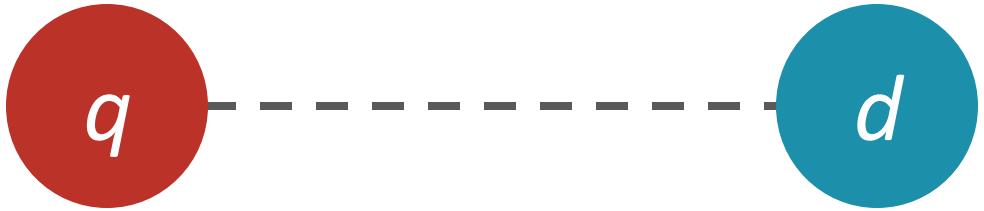
- Relevance of a document impacts other documents

The ranking problem



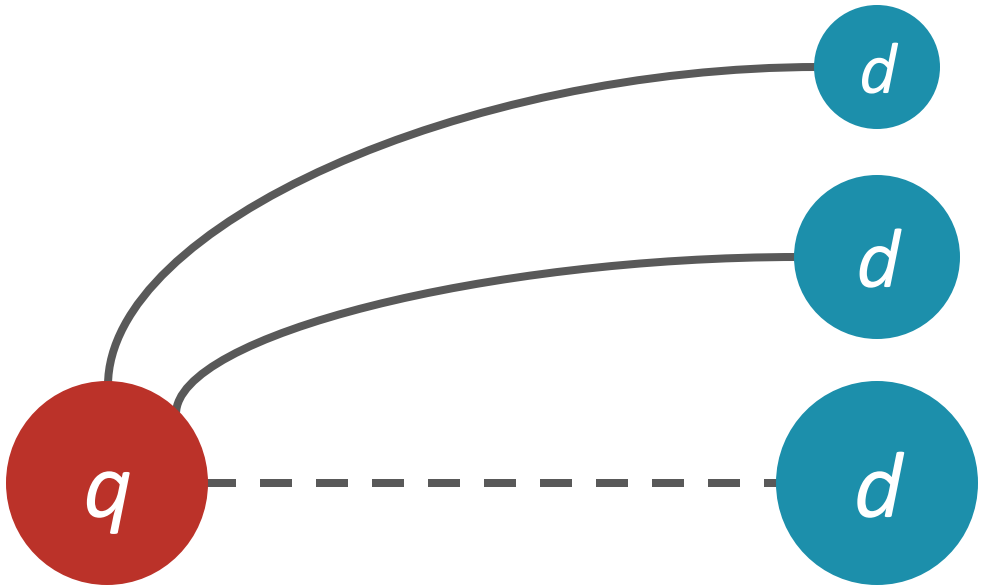
$$f(q, d)$$

The ranking problem



$$f(q, d, D)$$

The diversification problem



$$f(q, d, D)$$

How to compute $f(q,d,D)$?

Diversification dimensions

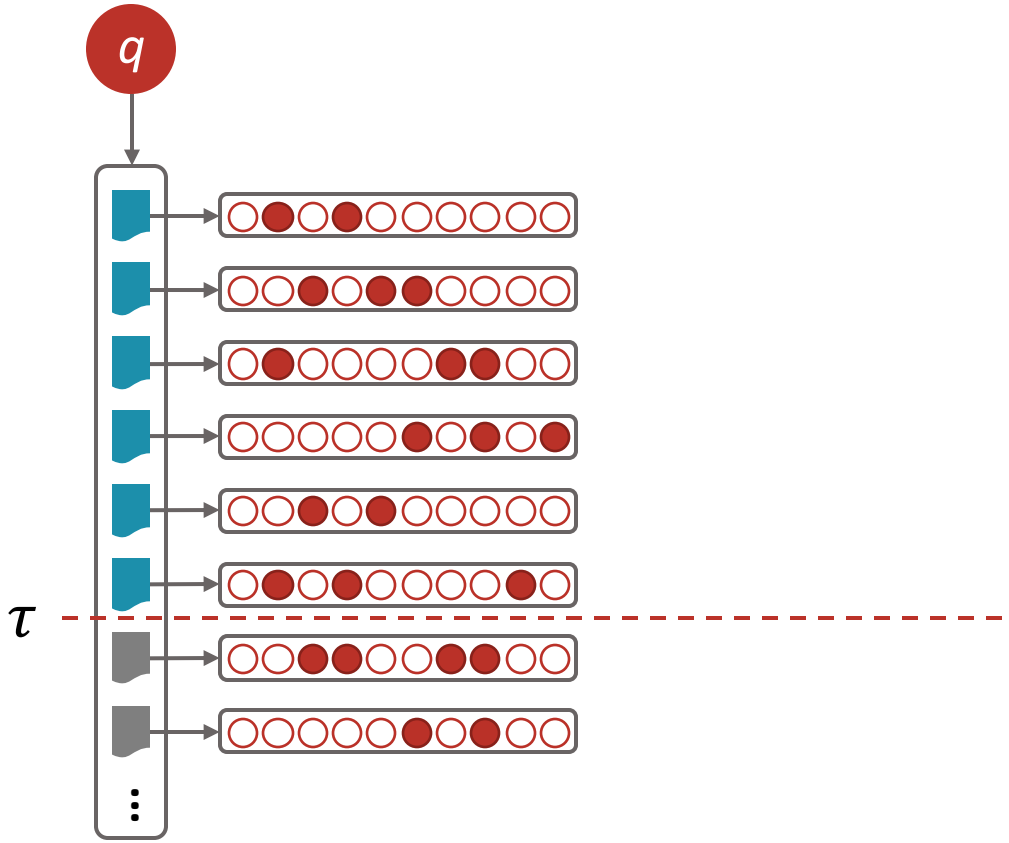
Aspect representation

- Implicit aspects
- Explicit aspects

Ranking strategy

- Novelty-based
- Coverage-based

Implicit aspects



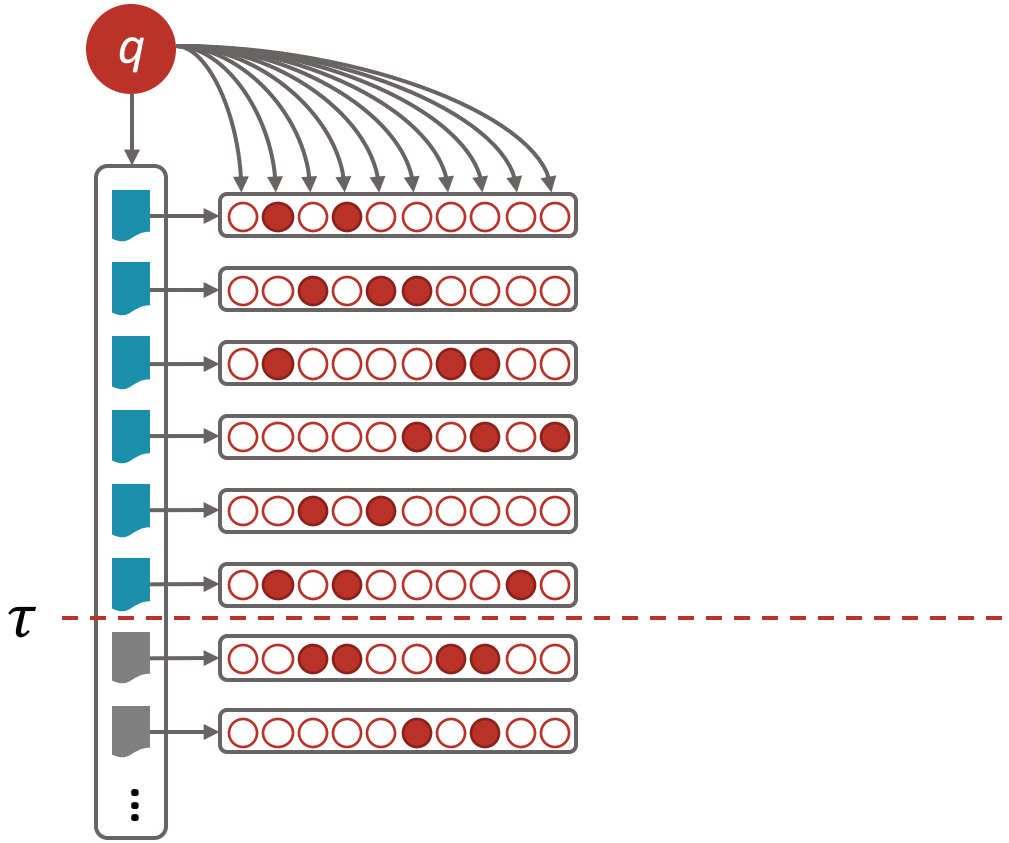
Raw factors

- One-hot term vectors
- Count-based term vectors

Latent factors

- Topic models
- Cluster models
- Embeddings

Explicit aspects



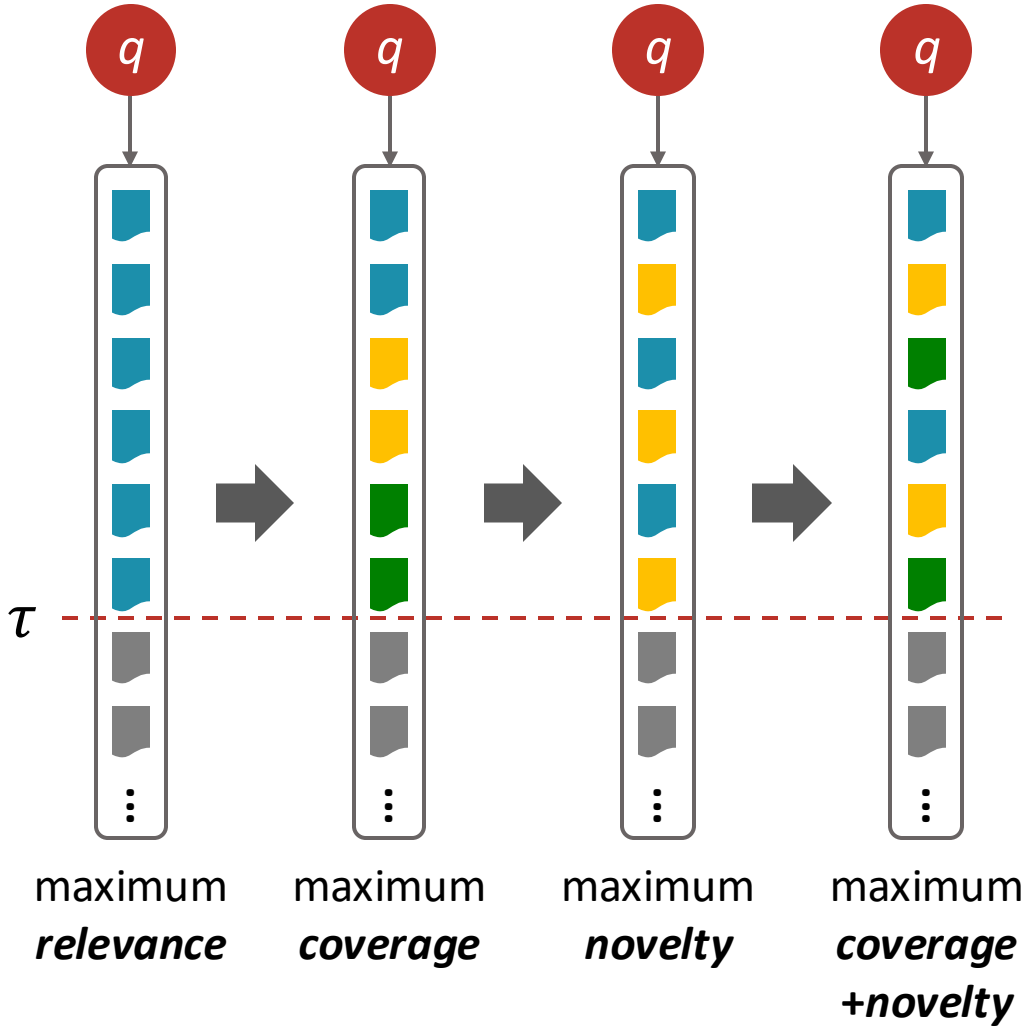
Fixed aspects

- Query categories

Non-fixed aspects

- Query suggestions

Ranking strategies



Coverage-based strategy

- Cover as many aspects as possible

Novelty-based strategy

- Cover aspects as early as possible

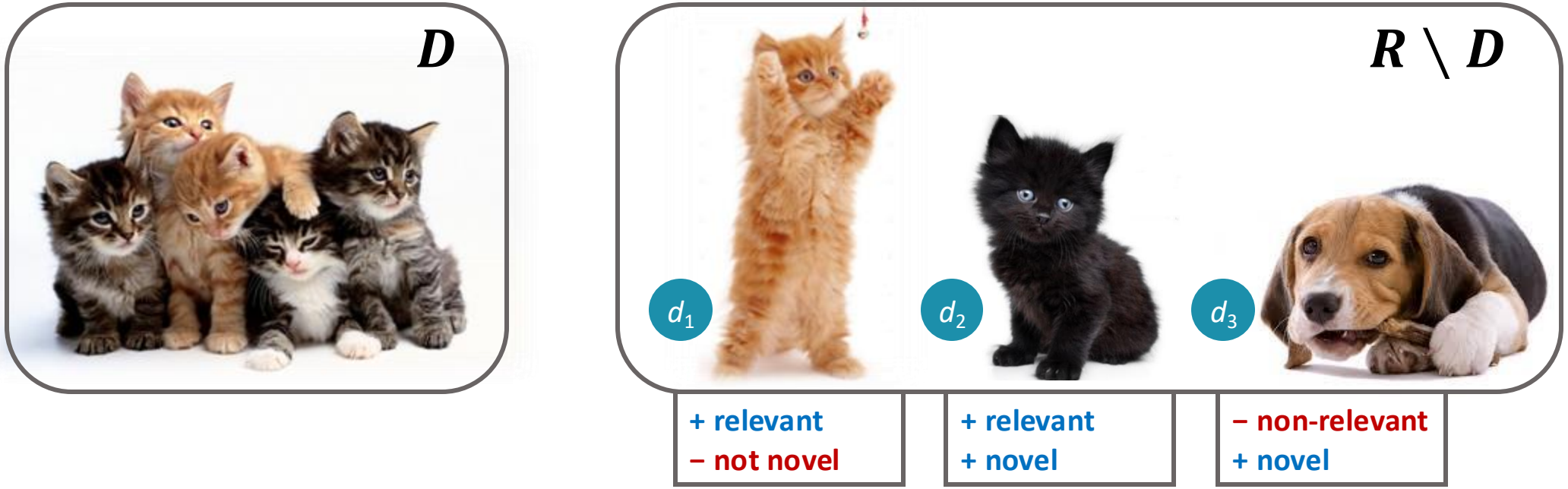
Hybrid strategy

- Promote coverage and novelty

MMR (Carbonell and Goldstein, 1998)

Maximal Marginal Relevance

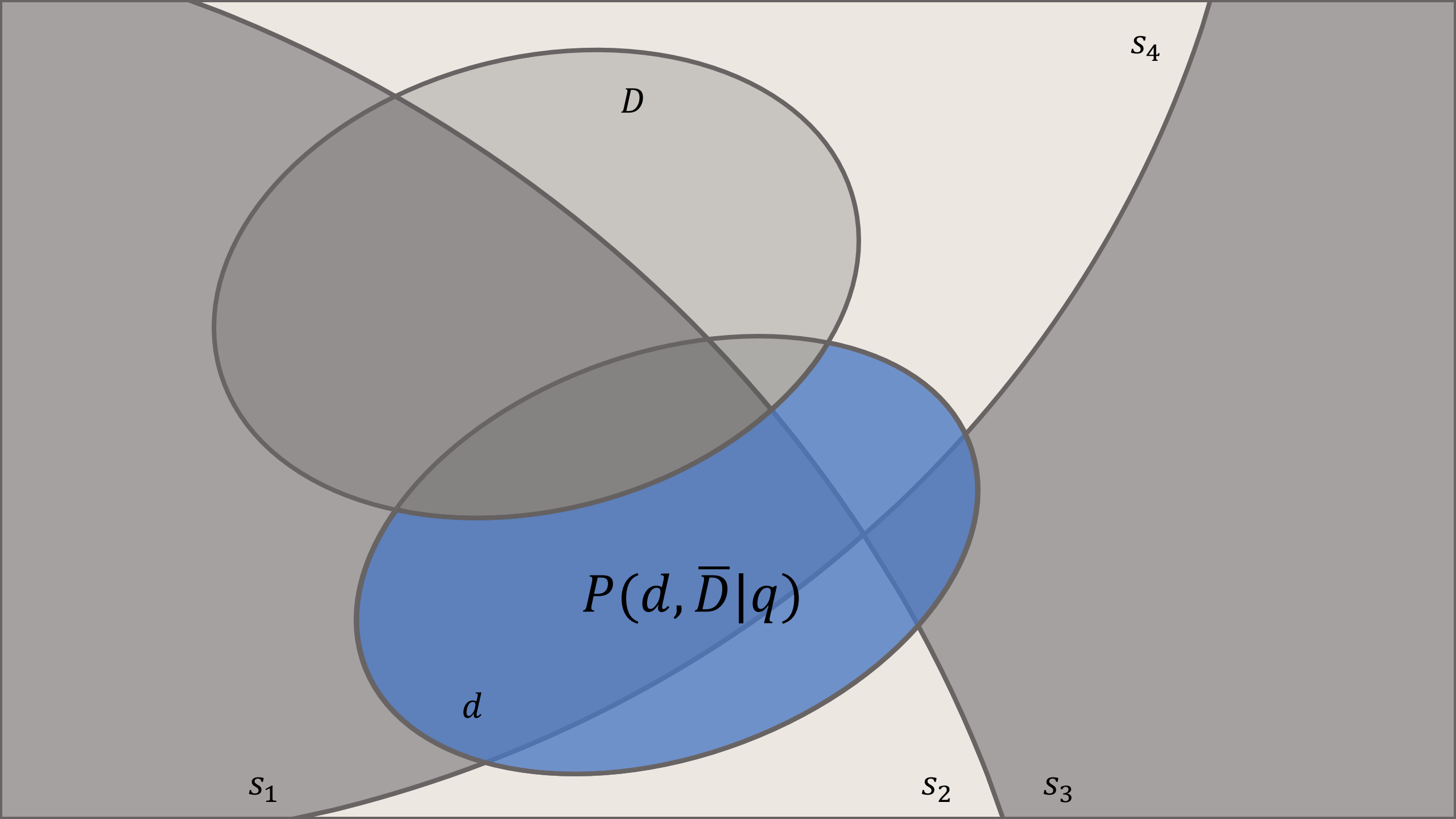
- $f(q, d, D) = \lambda\, rel(q, d) - (1 - \lambda) \max_{d_j \in D} sim(d, d_j)$

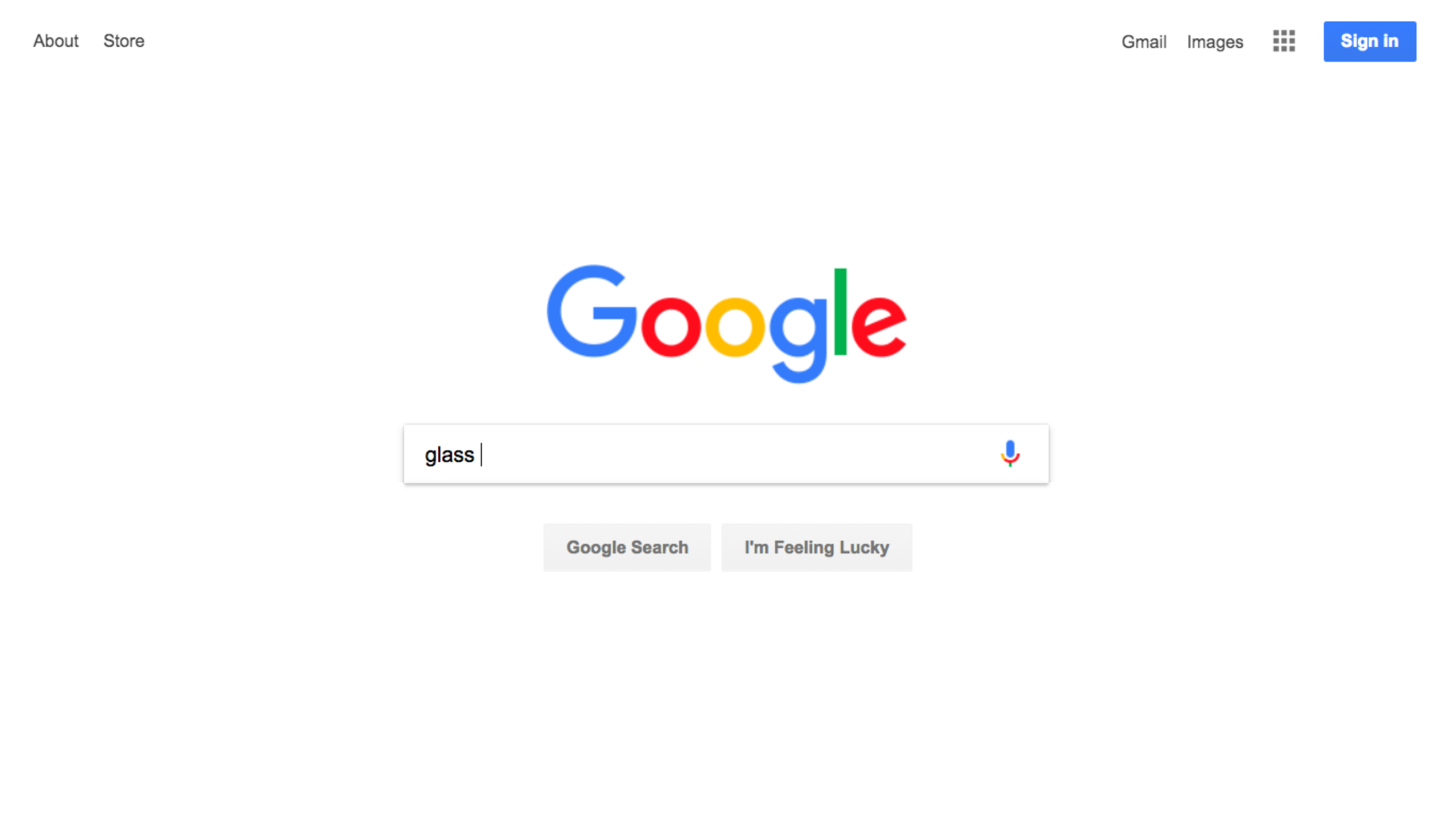


xQuAD (Santos et al., 2010)

Explicit Query Aspect Diversification

- $f(q, d, D) = (1 - \lambda)\ P(d|q) + \lambda\ P(d, \overline{D}|q)$


s_4
D
$P(d, \overline{D}|q)$
d
s_1
s_2
s_3


About
Store
Gmail
Images
Sign in
Google
glass
Google Search
I'm Feeling Lucky

glass

glass **door**
glass **castle**
glass **blunt**
glass **animals**
glass **house**
glass **house** – Museum in New Canaan, Connecticut
glass **house** – Concert hall in Pomona, California
glass **castle trailer**
glass **animals tour**
glass **repair**
glass **pipes**
glass **shower doors**

Google Search I'm Feeling Lucky

Report inappropriate predictions

Estimating $\boldsymbol{P(d, \bar{D}|q)}$

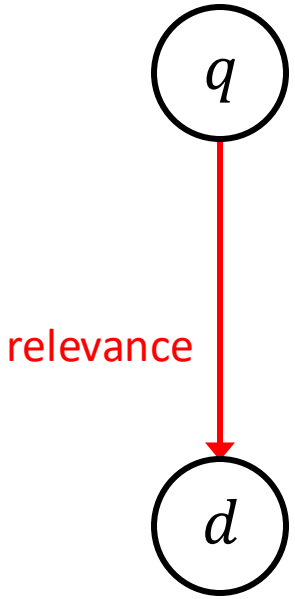
- $f(q, d, D) = (1 - \lambda)\, P(d|q)$

  $+ \lambda \sum_{s \in S} P(s|q) P(d|q, s) \prod_{d_j \in D} (1 - P(d_j|q, s))$

Estimating $P(d, \bar{D}|q)$

- $f(q, d, D) = (1 - \lambda)\, P(d|q)$

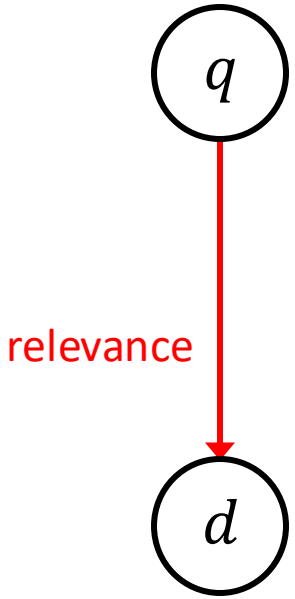
  $+ \lambda \sum_{s \in S} P(s|q) P(d|q, s) \prod_{d_j \in D} (1 - P(d_j|q, s))$

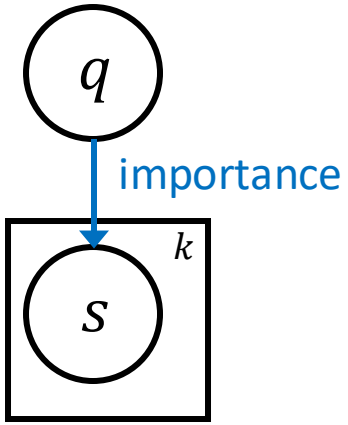


Estimating $P(d, \bar{D}|q)$

- $f(q, d, D) = (1 - \lambda)\, P(d|q)$

  $+ \lambda \sum_{s \in S} P(s|q) P(d|q, s) \prod_{d_j \in D} (1 - P(d_j|q, s))$

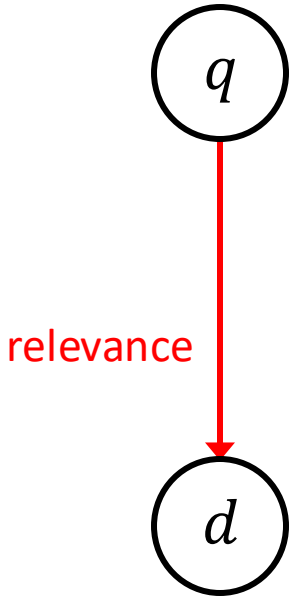


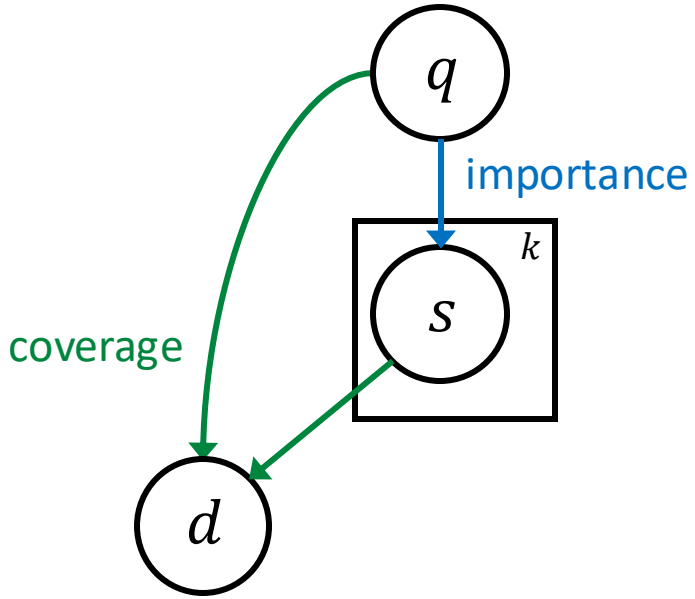


Estimating $P(d, \bar{D}|q)$

- $f(q, d, D) = (1 - \lambda)\, P(d|q)$

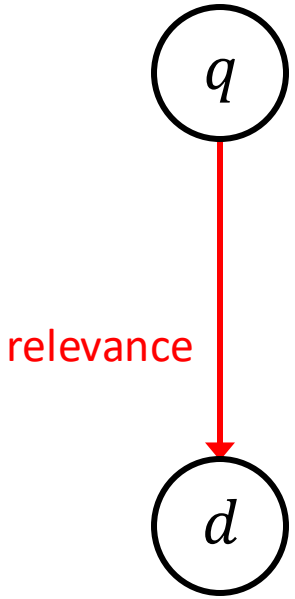
  $+ \lambda \sum_{s \in S} P(s|q) P(d|q, s) \prod_{d_j \in D} (1 - P(d_j|q, s))$

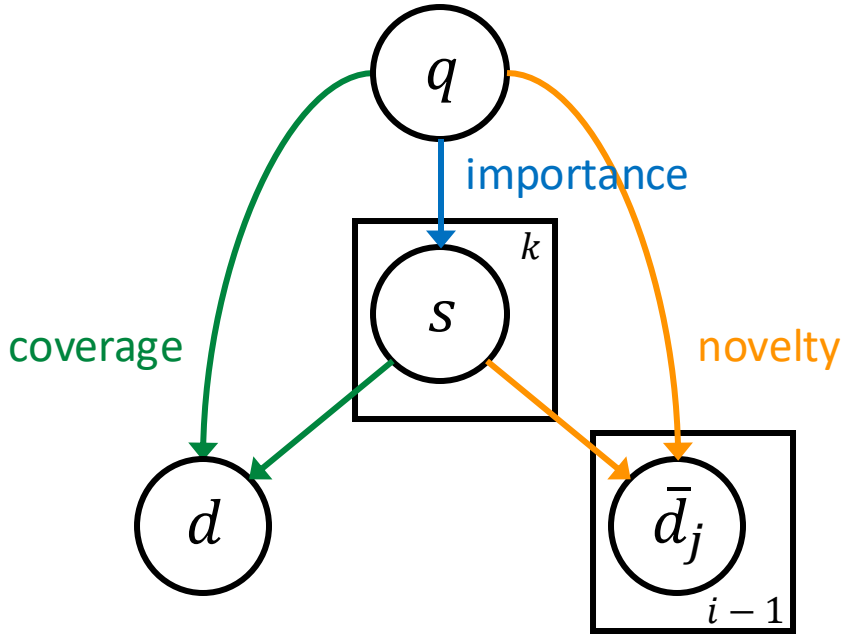




Estimating $P(d, \bar{D}|q)$

- $f(q, d, D) = (1 - \lambda)\, P(d|q)$
  $+ \lambda \sum_{s \in S} P(s|q) P(d|q, s) \prod_{d_j \in D} (1 - P(d_j|q, s))$





Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{D}^{(i-1)}$$
$$= (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

How to estimate $\mathbf{R}$, $\mathbf{C}$, $\mathbf{N}$, and λ?

- Advanced topics (see Santos et al., FnTIR 2015, Ch. 6)

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{D}^{(i-1)}$$
$$= (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{R} = \begin{bmatrix} 0.70 \\ 0.50 \\ 0.30 \\ 0.20 \\ 0.10 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 0.30 & 0.40 \\ 0.70 & 0.60 \\ 0.20 & 0.30 \\ 0.70 & 0.80 \\ 0.40 & 0.20 \end{bmatrix}, \mathbf{N}^{(0)} = \mathbf{I} = \begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix}$$

$$\lambda = 0.5$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(1)} = (1-0.5)\begin{bmatrix}0.70\\0.50\\0.30\\0.20\\0.10\end{bmatrix} + 0.5\begin{bmatrix}0.30 & 0.40\\0.70 & 0.60\\0.20 & 0.30\\0.70 & 0.80\\0.40 & 0.20\end{bmatrix}\begin{bmatrix}0.60\\0.40\end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(1)} = \begin{bmatrix} 0.52 \\ 0.58 \\ 0.26 \\ 0.47 \\ 0.21 \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix}$$

Example application

$$\mathbf{N}^{(i)} = \mathrm{diag}(\mathbf{1} - \mathbf{C}_r)\mathbf{N}^{(i-1)}$$

$$\mathbf{C} = \begin{bmatrix} 0.30 & 0.40 \\ 0.70 & 0.60 \\ 0.20 & 0.30 \\ 0.70 & 0.80 \\ 0.40 & 0.20 \end{bmatrix}, \mathbf{N}^{(0)} = \begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix}$$

Example application

$$\mathbf{N}^{(i)} = \mathrm{diag}(\mathbf{1} - \mathbf{C}_r)\mathbf{N}^{(i-1)}$$

$$\mathbf{N}^{(1)} = \begin{bmatrix} 1.00 - 0.70 & 0.00 \\ 0.00 & 1.00 - 0.60 \end{bmatrix} \begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix}$$

$$= \begin{bmatrix} 0.30 & 0.00 \\ 0.00 & 0.40 \end{bmatrix} \begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix}$$

$$= \begin{bmatrix} 0.18 \\ 0.16 \end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(2)} = (1-0.5)\begin{bmatrix}0.70\\0.30\\0.20\\0.10\end{bmatrix} + 0.5\begin{bmatrix}0.30 & 0.40\\0.20 & 0.30\\0.70 & 0.80\\0.40 & 0.20\end{bmatrix}\begin{bmatrix}0.18\\0.16\end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(2)} = \begin{bmatrix} 0.41 \\ 0.19 \\ 0.23 \\ 0.10 \end{bmatrix} \begin{matrix} d_1 \\ d_3 \\ d_4 \\ d_5 \end{matrix}, \mathbf{N}^{(2)} = \begin{bmatrix} 0.13 \\ 0.10 \end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(3)} = (1-0.5)\begin{bmatrix}0.30\\0.20\\0.10\end{bmatrix} + 0.5\begin{bmatrix}0.20 & 0.30\\0.70 & 0.80\\0.40 & 0.20\end{bmatrix}\begin{bmatrix}0.13\\0.10\end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(3)} = \begin{bmatrix} 0.17 \\ 0.18 \\ 0.08 \end{bmatrix} \begin{matrix} d_3 \\ d_4 \\ d_5 \end{matrix}, \mathbf{N}^{(3)} = \begin{bmatrix} 0.04 \\ 0.02 \end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(4)} = (1-0.5)\begin{bmatrix}0.30\\0.10\end{bmatrix} + 0.5\begin{bmatrix}0.20 & 0.30\\0.40 & 0.20\end{bmatrix}\begin{bmatrix}0.04\\0.02\end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1-\lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(4)} = \begin{bmatrix} 0.16 & d_3 \\ 0.06 & d_5 \end{bmatrix}, \mathbf{N}^{(4)} = \begin{bmatrix} 0.03 \\ 0.01 \end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1 - \lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(5)} = (1 - 0.5)[0.10] + 0.5\,[0.40 \quad 0.20]\begin{bmatrix}0.03\\0.01\end{bmatrix}$$

Example application

$$\mathbf{X}^{(i)} = (1 - \lambda)\,\mathbf{R} + \lambda\,\mathbf{CN}^{(i-1)}$$

$$\mathbf{X}^{(5)} = [0.06]\,d_5$$

Example application

Input ranking

- $R = (d_1, d_2, d_3, d_4, d_5)$

Diversified ranking

- $D = (d_2, d_1, d_4, d_3, d_5)$

Summary

Query ambiguity can harm search quality

- Search result diversification can help

NP-hard problem

- Efficient greedy approximation

Several objective functions in the literature

- Explicit/hybrid approaches are the state-of-the-art

Open directions

Aspect representation

- Exploitation of dependencies between aspects
- Document-driven aspect identification

Diversification strategy

- Personalized diversification
- Discriminative diversification

References

The use of MMR, diversity-based reranking for reordering documents and producing summaries

Carbonell and Goldstein, SIGIR 1998

Exploiting query reformulations for Web search result diversification

Santos et al., WWW 2010

References

Search result diversification

Santos et al., FnTIR 2015

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

Learning to Rank: Fundamentals

Rodrygo L. T. Santos

rodrygo@dcc.ufmg.br