

Machine Learning Approaches to Information Retrieval

Hang Li

Microsoft Research Asia

Joint Work with Jun Xu, Yunbo Cao, Guoping Hu

Talk Outline

- Introduction to Information Retrieval
- Search by Type
- Factoid Search
- New Learning Algorithm for Ranking
- Information Desk
- Summary

Introduction to Information Retrieval

What Is Information Retrieval?



I want to access information X



query

information

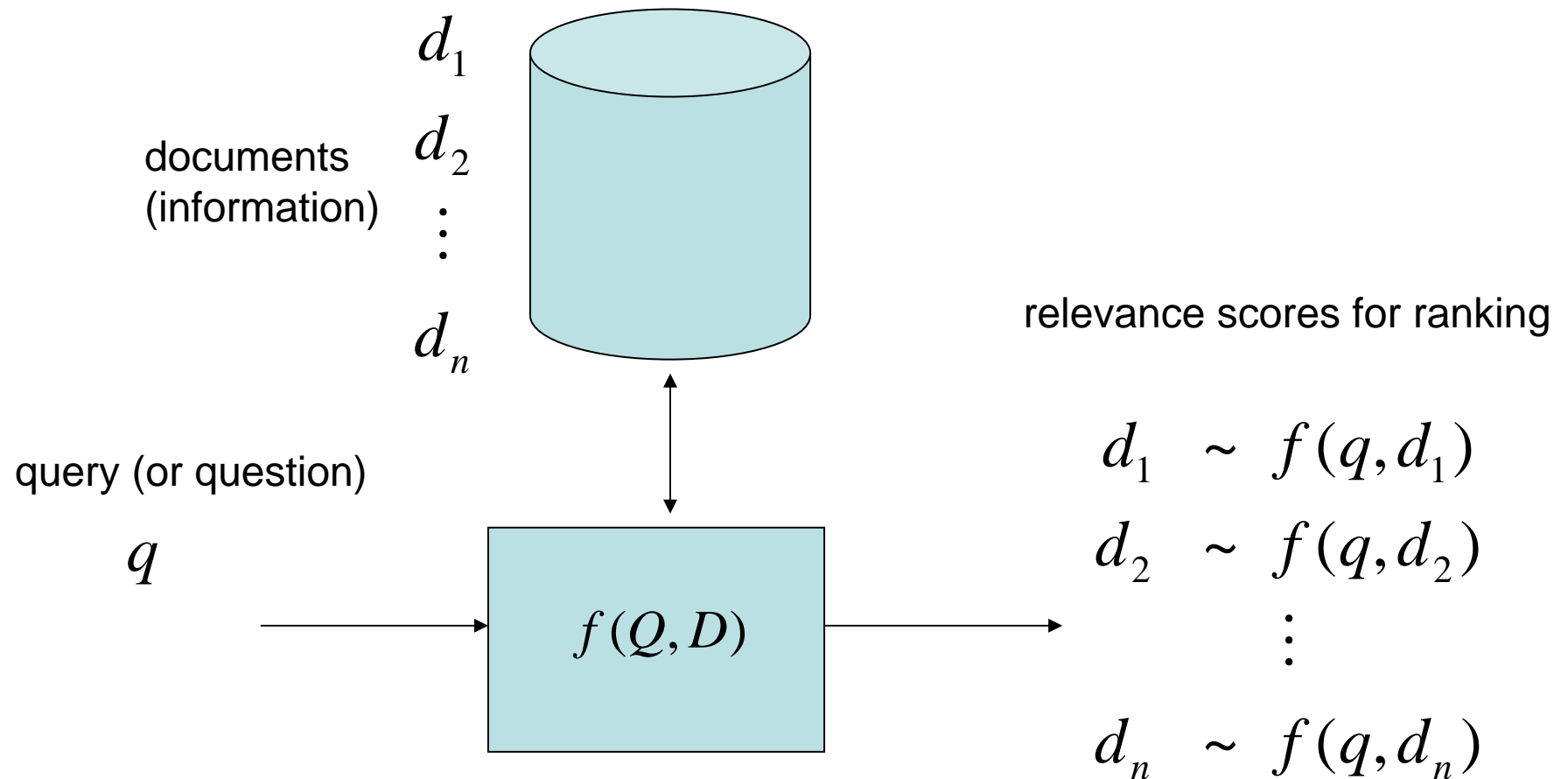


Why Important?

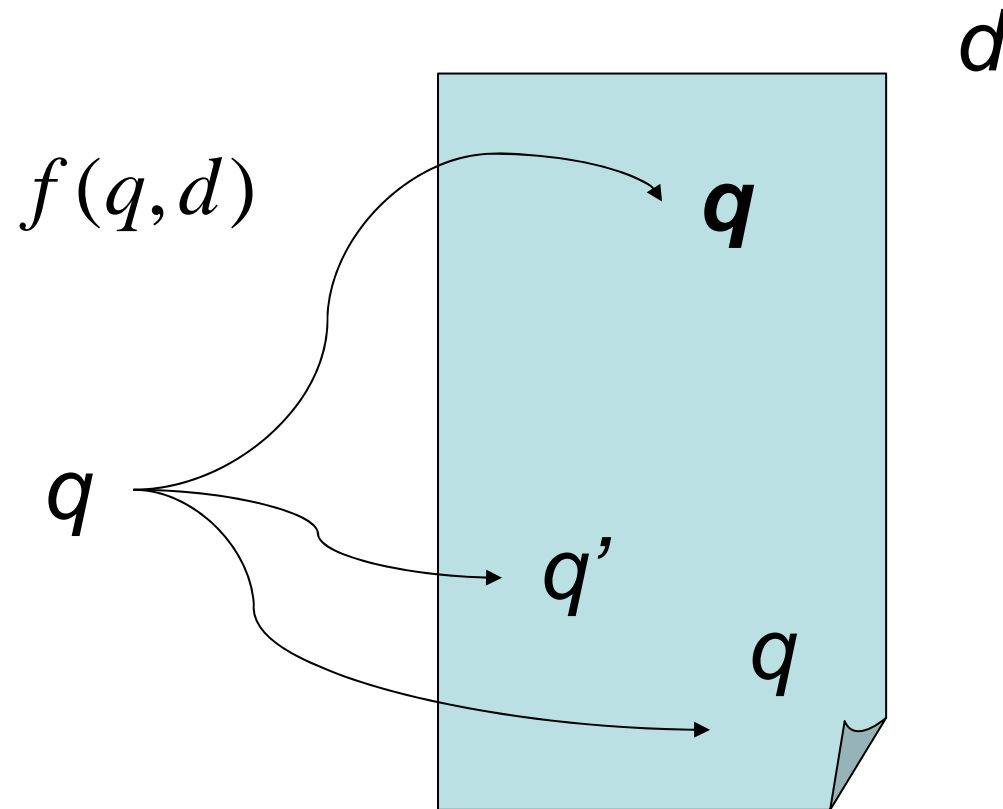


Current Approach
Information Retrieval = Ranking

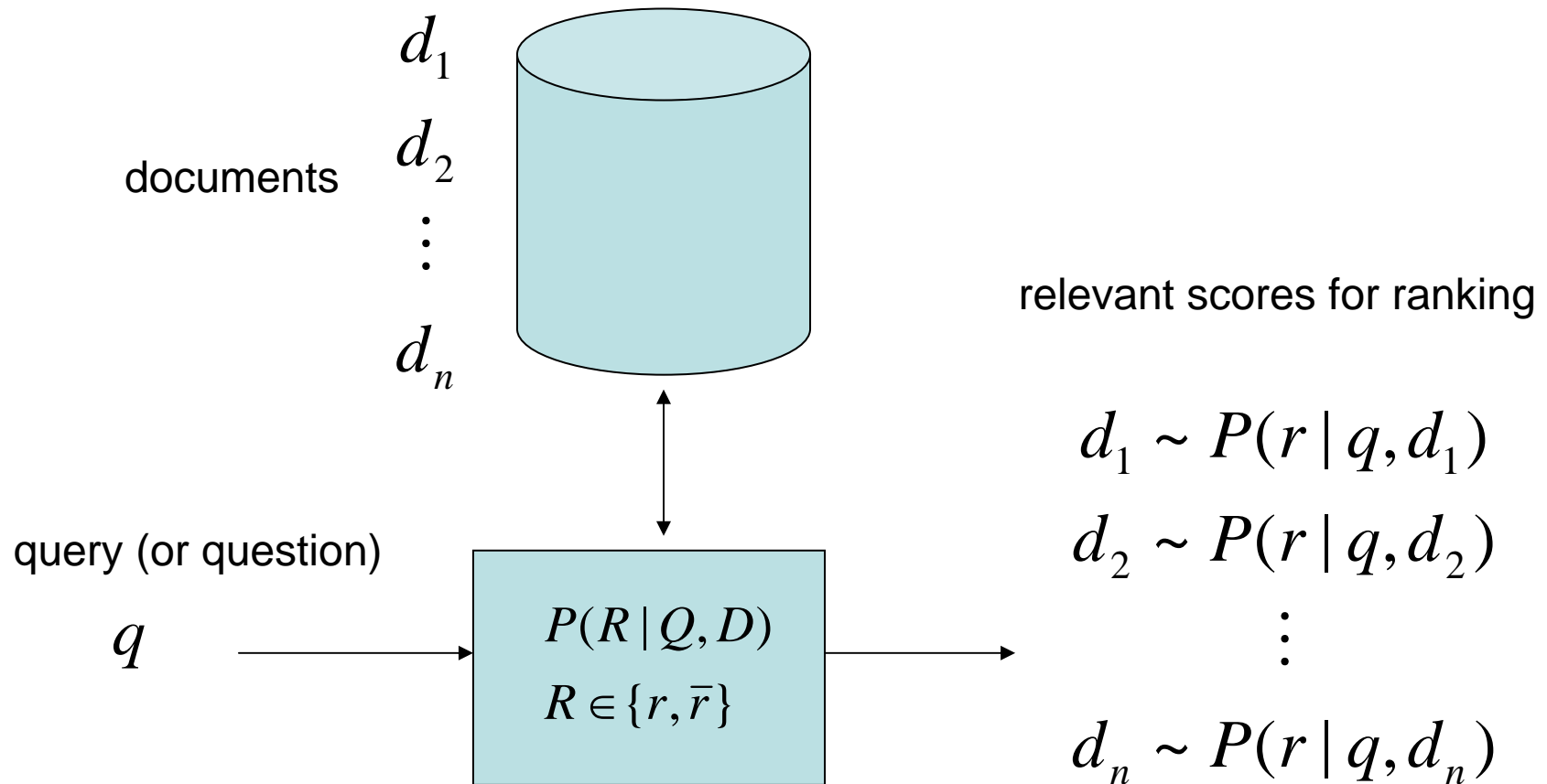
General Model for Ranking



Relevance: Matching between Query and Document

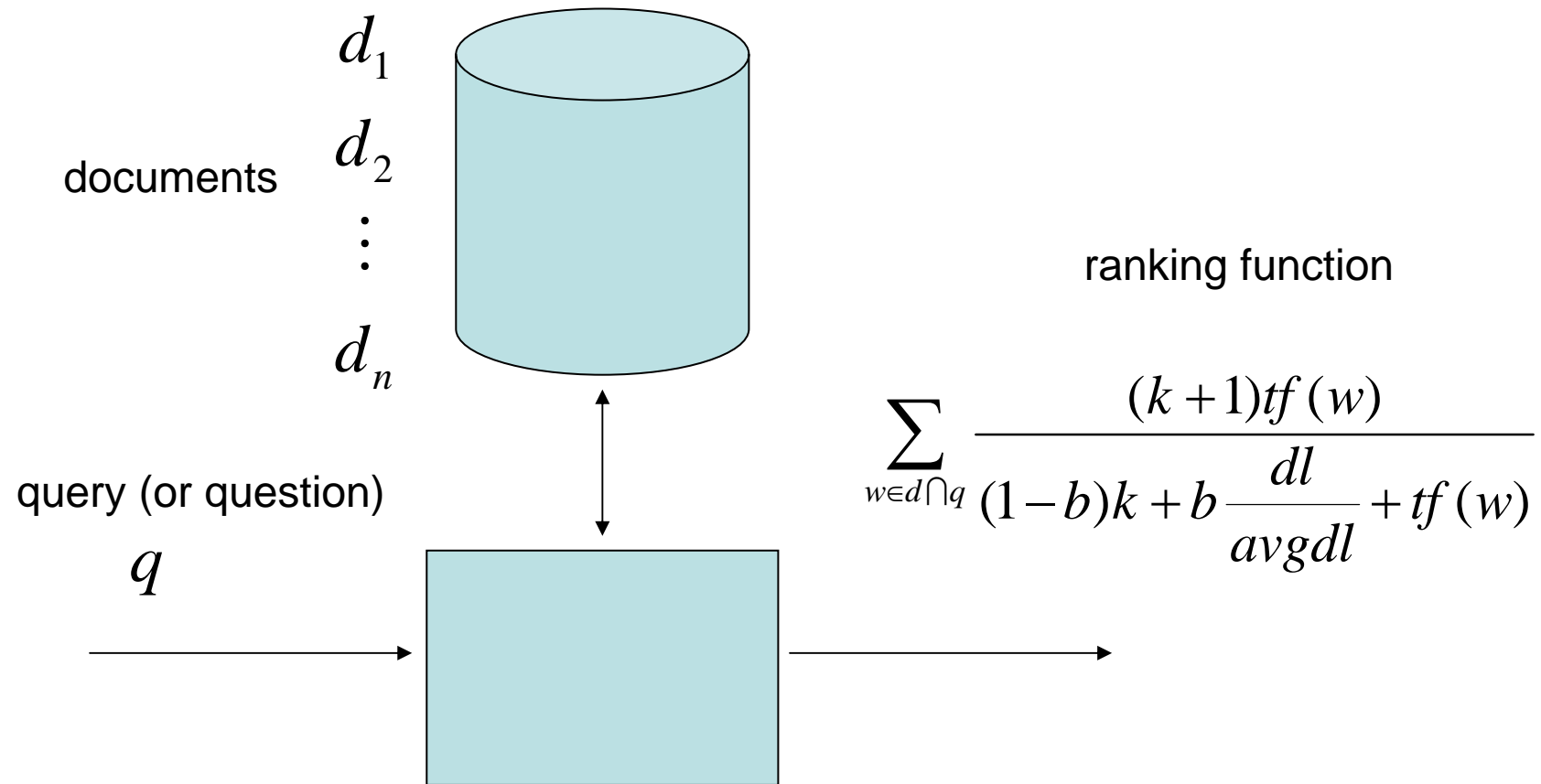


Probabilistic Model



Okapi or BM25

(Robertson and Walker 1994)



Language Mode

(Ponte and Croft 1998; Lafferty and Zhai, 2001)

document = bag of words

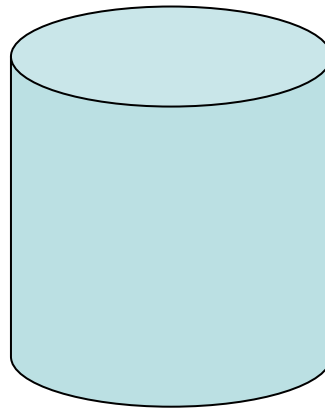
$$d_1 = w_{11} w_{12} \cdots w_{1l_1}$$

$$d_2 = w_{21} w_{22} \cdots w_{2l_2}$$

⋮

$$d_n = w_{n1} w_{n2} \cdots w_{nl_n}$$

$$q = w_{q1} w_{q2} \cdots w_{ql_q}$$



relevance scores for ranking

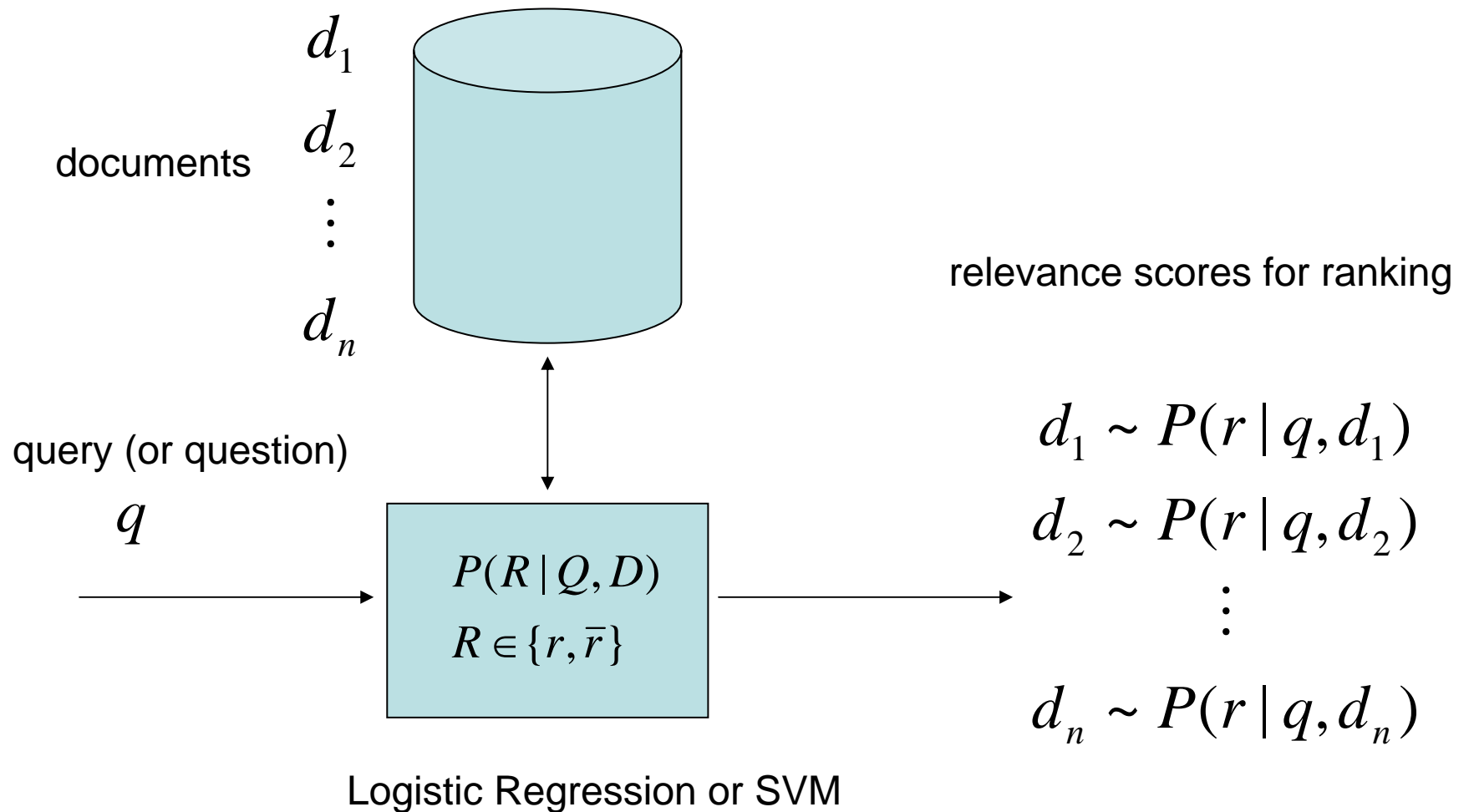
$$d_1 \sim P(q | d_1)$$

$$d_2 \sim P(q | d_2)$$

⋮

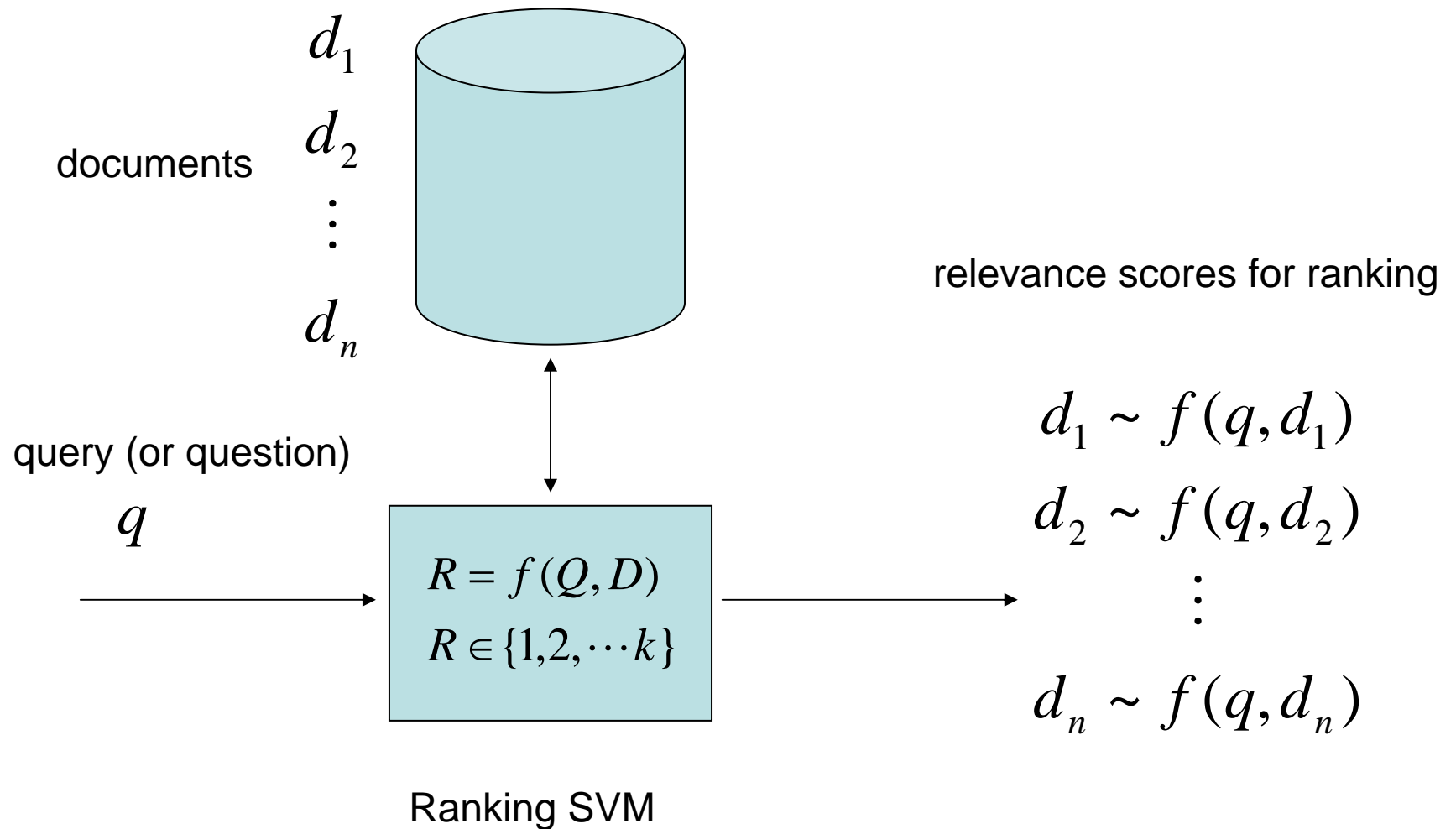
$$d_n \sim P(q | d_n)$$

Classification Model

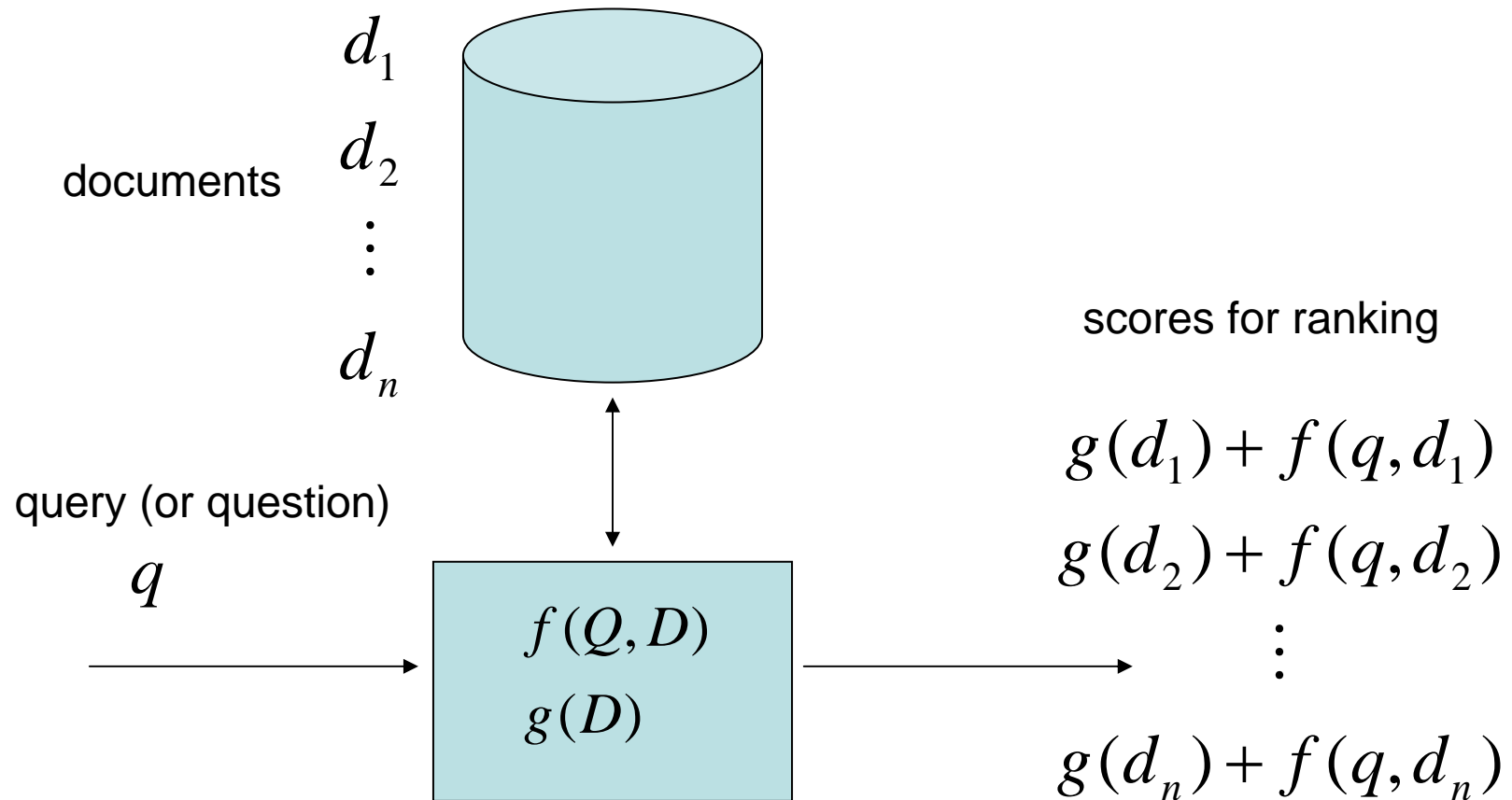


Ordinal Regression Model

(Herbrich et al., 2000; Joachims, 2002)

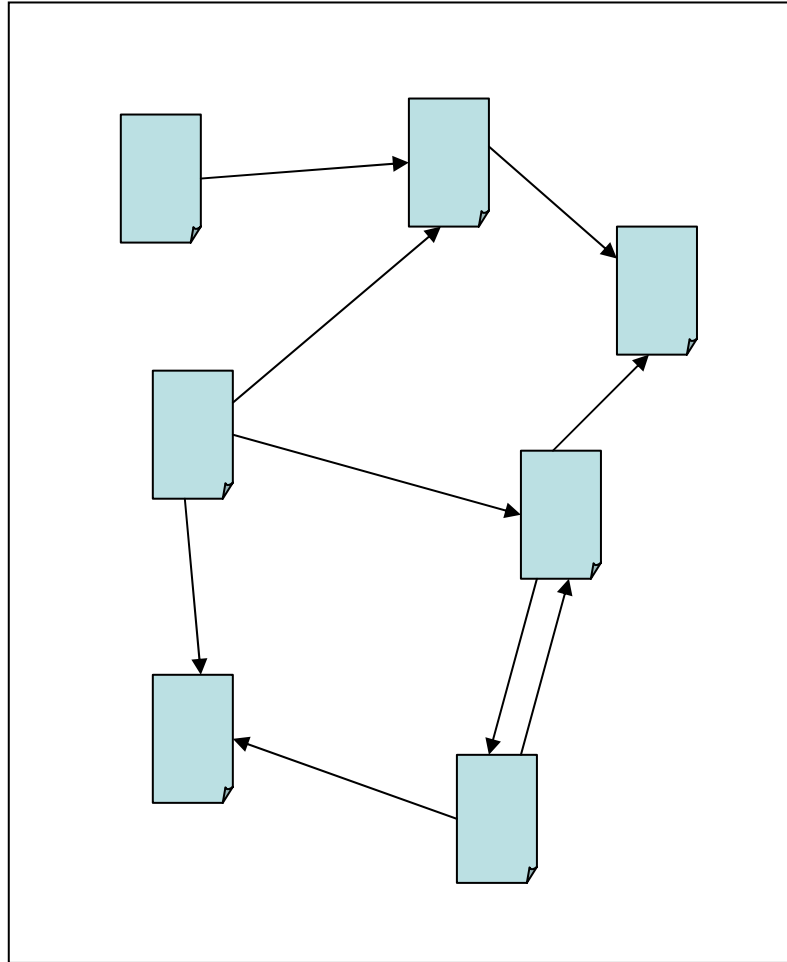


General Model for Ranking (2)



Page Rank

(Brin and Page, 1998)



$$P(d_i) = \alpha \frac{1}{n} + (1 - \alpha) \sum_{d_j \in M(d_i)} \frac{P(d_j)}{L(d_j)}$$

Challenges

- Proximity
- Synonym and polysemi
- Quality and freshness of document (webpage)
- Spamming
- Information granularity
- Search need understanding
- Evaluation
- Personalization
- Training data collection (relevance feedback, click-through)

Search by Document Type

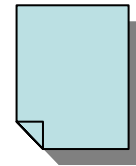
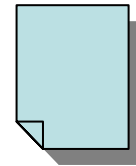
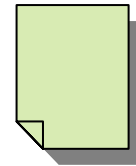
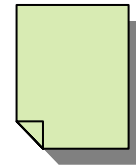
Search by Document Type

I want to find documents relevant to X in type Y

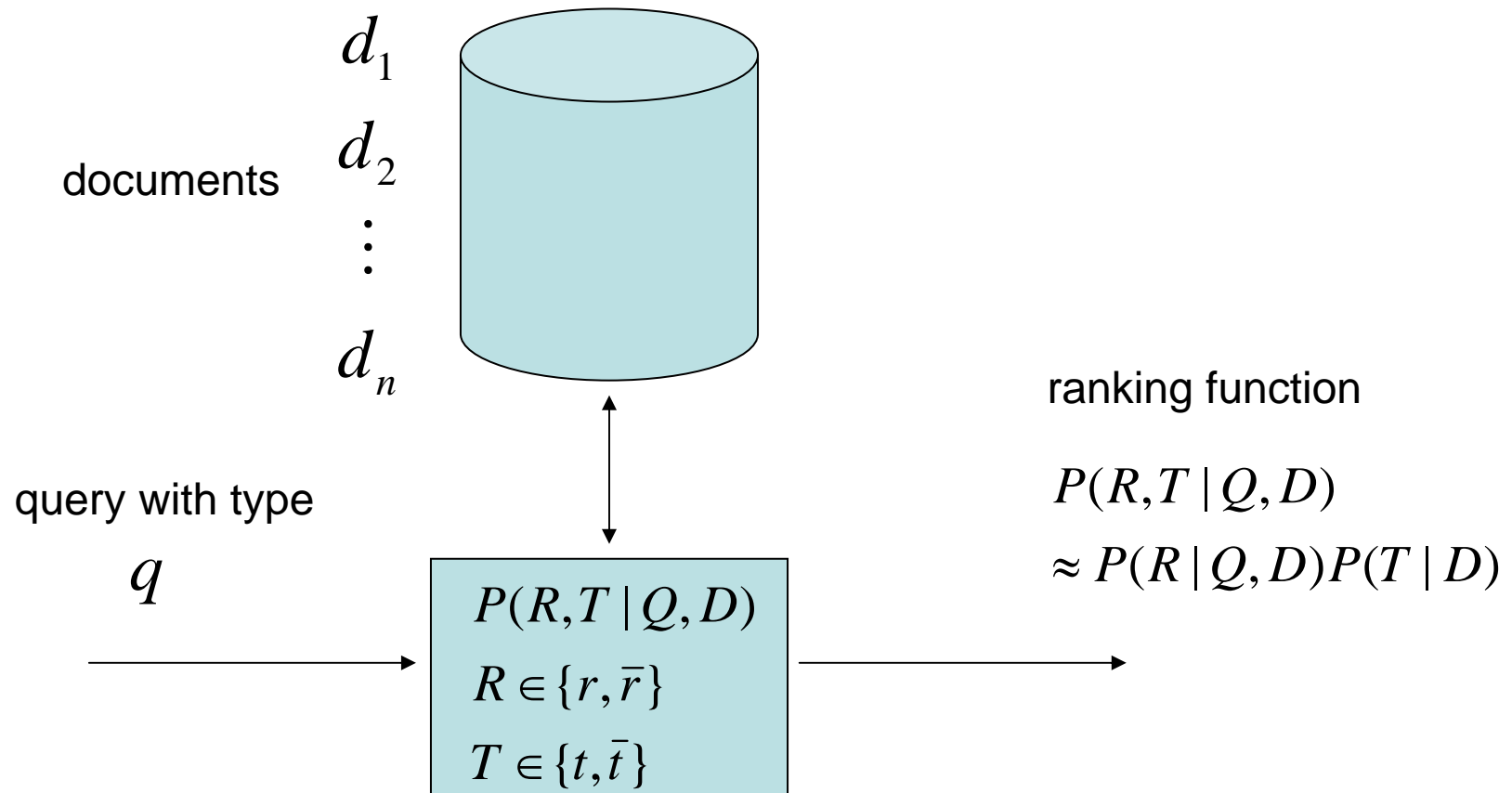


query

documents



Probabilistic Model



Example – Manuals Search

- Query
“create an external link”
- Document1
 - External Links**

External links are to Web sites external to your Web site. You can create external links by using buttons, highlighting text, or creating hot spots on images.

For example creating a linkage to the Far Eastern Economic Review [FEER] one would go to the FEER through the Internet. You might want to right click on the FEER logo or recent issue cover and save the image to your computer [must be saved as a graphical interface file - GIF]. To make either the button or the text or the picture (or all three) interactive you highlight what you want to connect and then you click the hyperlink tool, or the Edit-Hyperlink. Press the World Wide Web tab in the Create Hyperlink dialogue box [see box below]. Whatever address you have in the running Net browser will show up in the box, or you can browse. In this case, the FEER address is there, one only need to press OK and the external hyperlink is created. All three examples to the right will link you to the FEER site.
- Document2
 - Creating an External Link**

To create an anchor that is a link to another document:
Select to select (by click and drag or by keyboard) the text for the link you are creating.
Click the Link button (first case) or select the entry "Create or change link" in the Links menu (second case).
In the first case, the cursor changes from an arrow to a hand to let you click the target document.
If the target document is displayed in another Amaya window, click anywhere within that window to create the link.
If the target document is not displayed in another Amaya window, press the F2 or Delete key, or click a part of the document which cannot be a valid target. A dialog prompts you for the location of the target document. Type the URI of the target document and then Confirm to create the link.
In the second case, a dialog prompts you for the location of the target document.
If the target document is displayed in another Amaya window and you want to select it by click, click the Click button then click anywhere within that window to create the link.
If the target document is not displayed in another Amaya window, type the URI of the target document and then Confirm to create the link.

Manuals Search -- Experiment

- MS intranet data
 - 50 queries from log of Microsoft Web
 - 1.22 answers per query (from 5000 documents)

- Evaluation Measure

$$RR_i = \frac{1}{Rank_i}$$

$$MRR = \frac{\sum_{i=1}^Q RR_i}{Q}$$

- Results

Method	MRR
Type Only	0.3651
Relevance Only	0.5688
Combined Model (our approach)	0.7278

Factoid Search

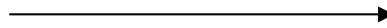
People Search

(Cao et al., 2005)

I want to find people related to X



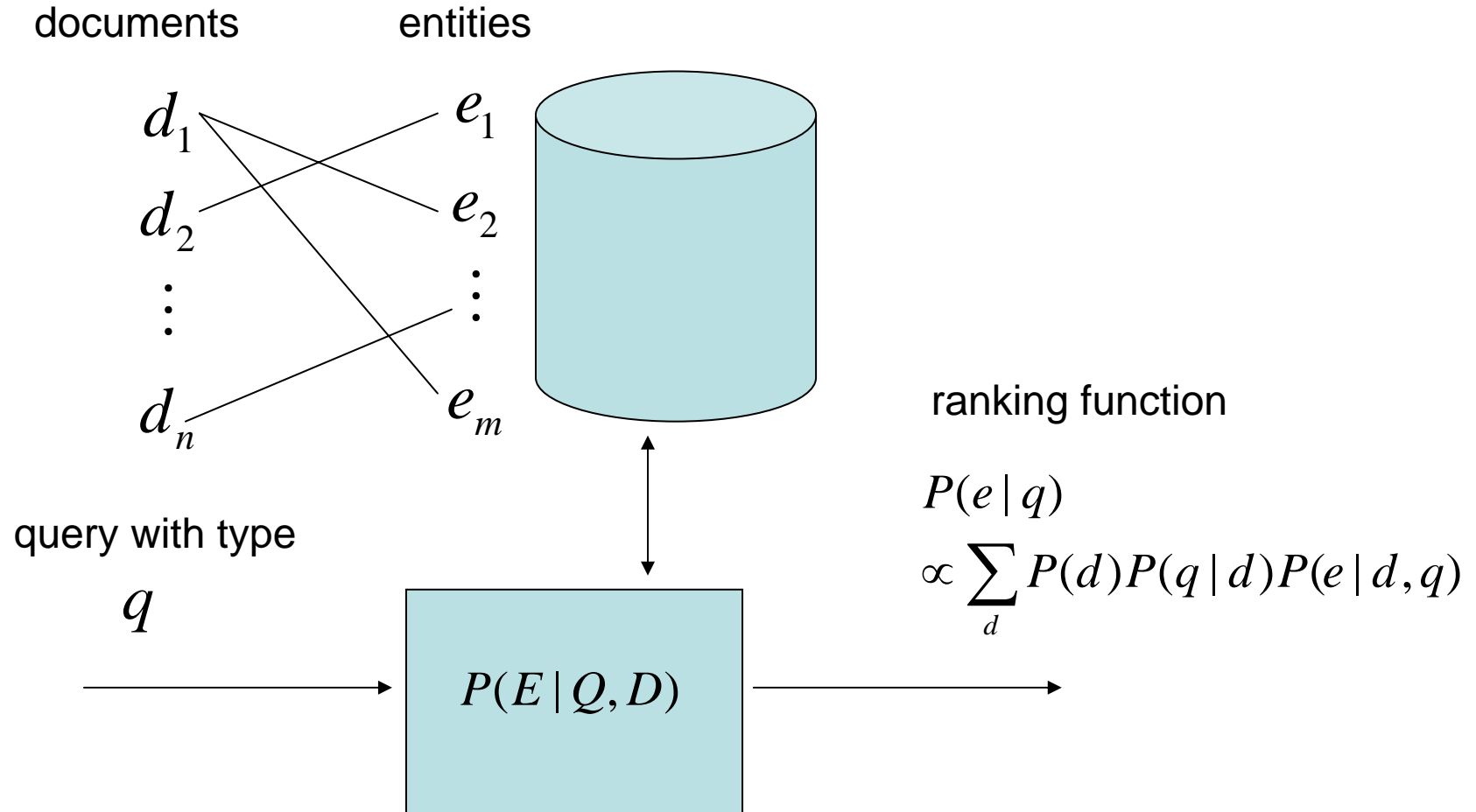
query



people



Probabilistic Model



Example – People Search

- Query
“Who knows about digital ink”
- Document
Jian Wang, Ph.D.
Senior Researcher & Research Manager
Multimodal User Interface Group, Microsoft Research Asia
Dr. Jian Wang is Research Manager of the Multimodal User Interface Group at Microsoft Research Asia (MSR Asia). Dr. Jian Wang's research specializations are ink and pen computing, usability, multimodal user interface, virtual reality and human cognition.
The Multimodal UI Group's current research projects include: advanced **digital ink** parser, digital ink annotation and representation of digital ink for Tablet PC. The group previously invented an inline input and correction user interface for Asian languages called Modeless Input User Interface, which allows Chinese users of Office XP to smoothly enter English and Chinese text without constantly switching between input language modes.
- Answer
Jian Wang

People Search -- Experiment

- MSR Corpus
 - 32 queries searching in 3109 documents
 - 810 Candidates
- Evaluation Measure

$$\text{Top-5 Precision} = \frac{\#\{\text{persons which appear in both Top-5 ranked candidates and ground truth}\}}{5}$$

- Results

Method	Top-5 Precision
Profile-based Model	0.428
Our Model	0.563 (+31.5%)

Time Search

I want to find time related to X

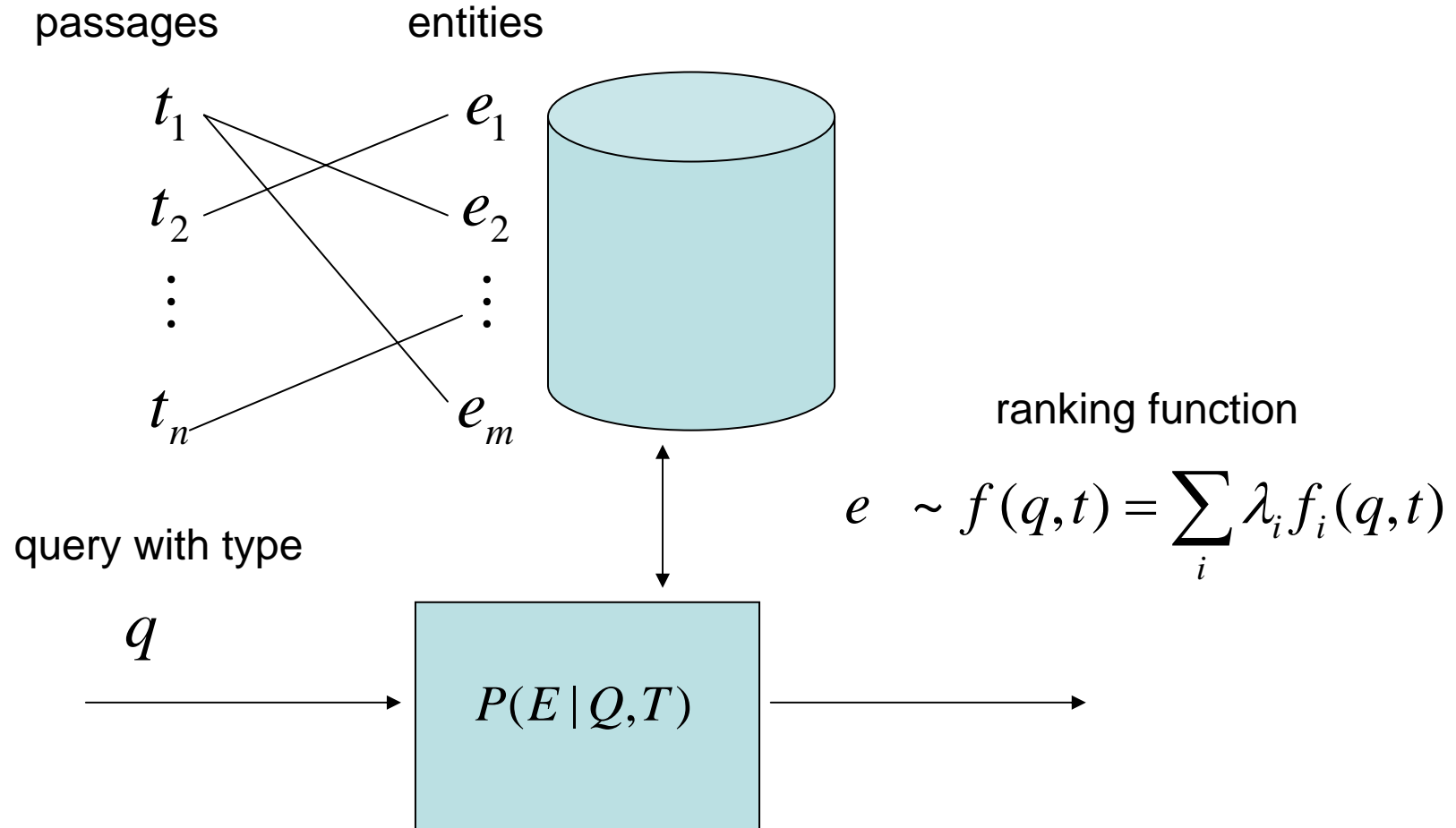


query →

← dates



Probabilistic Model



Time Search -- Example

- Query:
Nixon visit China
- Document:
Nixon's most significant achievement in foreign affairs may have been the establishment of direct relations with the People's Republic of China after a 21-year estrangement. Following a series of low-level diplomatic contacts in 1970 and the lifting of U.S. trade and travel restrictions the following year, the Chinese indicated that they would welcome high-level discussions, and Nixon sent his national security adviser, Henry Kissinger, to China for secret talks. The thaw in relations became apparent with the “ping-pong diplomacy” conducted by American and Chinese table-tennis teams in reciprocal visits in 1971–72. Nixon's visit to China in February–March 1972, the first by an American president while in office, concluded with the Shanghai Communiqué, in which the United States formally recognized the “one-China” principle—that there is only one China, and that Taiwan is a part of China.
- Answer
February–March 1972

Time Search -- Experiment

- MS intranet data
 - 100 queries from log of <http://msweb>, containing ‘when’, ‘schedule’, ‘day’, and ‘time’
 - 8.73 answers per query (from 10000 documents)

- Evaluation Measure

$$RR_i = \frac{1}{Rank_i}$$

$$MRR = \frac{\sum_{i=1}^Q RR_i}{Q}$$

- Results

Method	MRR
Best Baseline	0.5033
Learning	0.5809 (+15%)

New Learning Algorithm for Ranking

Ranking Learning

- Given:
 - $S = \{(\vec{x}_i, y_i)\}_{i=1}^m \subset X \times Y$, where $Y = \{r_1 \prec \dots \prec r_q\}$
 - $H = \{h: X \mapsto Y\}$ (hypothesis space)
 - $L: Y \times Y \mapsto R$ (loss function)
- Question: based on S find $h^* = \arg \min \mathbf{E}(L(h(\vec{x}), y))$
- ERM: choose $h_{ERM}^* = \operatorname{argmin} \sum_{i=1}^m L(h(\vec{x}_i), y_i)$
- SRM: choose $h_{SRM}^* = \operatorname{argmin} \sum_{i=1}^m L(h(\vec{x}_i), y_i) + \lambda Q(h)$

Viewing Ranking as Classification

- Formulating ranking problem as classification of example pairs
- Measuring the loss of h by inversions

$$L(y_1, y_2, \hat{y}_1, \hat{y}_2) = \begin{cases} 1 & (y_1 \prec y_2) \wedge \neg(\hat{y}_1 \prec \hat{y}_2) \\ 1 & (y_1 \succ y_2) \wedge \neg(\hat{y}_1 \succ \hat{y}_2) \\ 0 & \textit{otherwise} \end{cases}$$

Viewing Ranking as Classification (cont')

- Given $S = \{(\vec{x}_i, y_i)\}_{i=1}^m$, find h_{ERM} that minimizes

$$\sum_{i=1}^m \sum_{j=1}^m L(y_i, y_j, h(\vec{x}_i), h(\vec{x}_j))$$

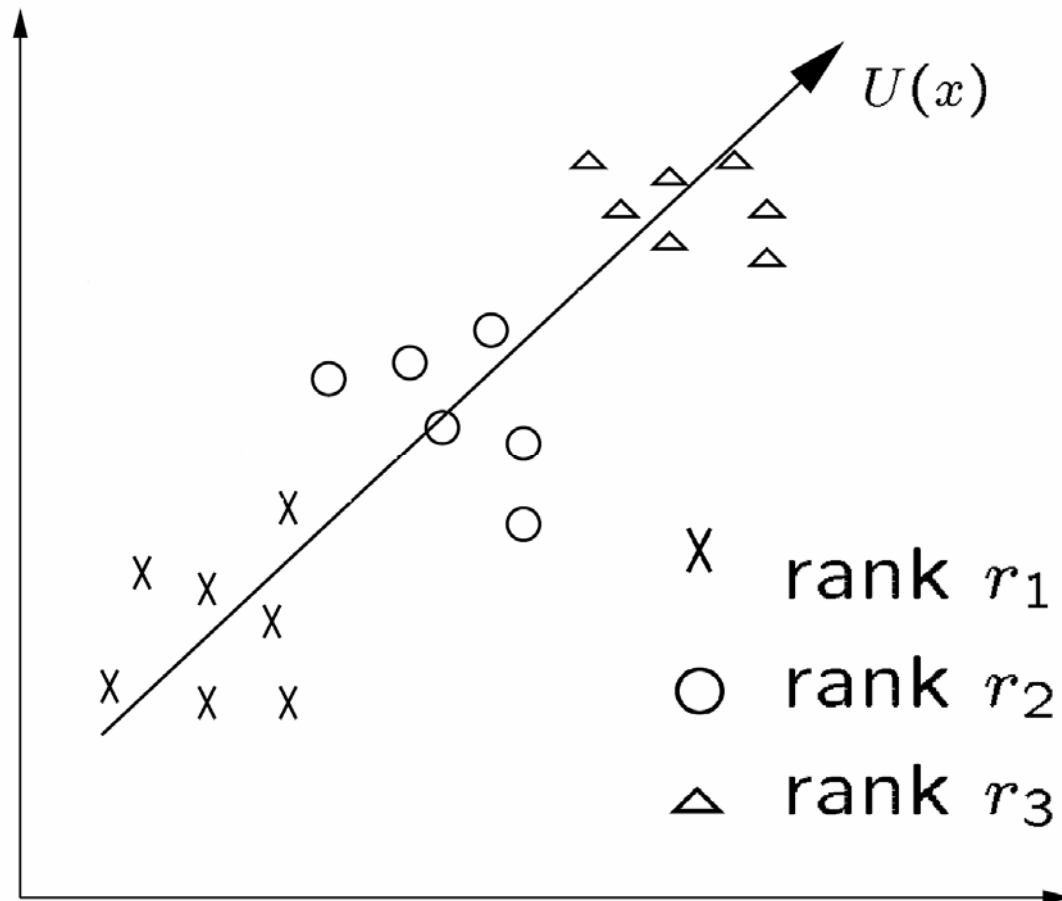
- Equivalently, find a f_h that minimizes

$$\sum_{((\vec{x}_i^{(1)}, \vec{x}_i^{(2)}), z_i) \in S'} L_{0-1}(f_h(\vec{x}_i^{(1)}, \vec{x}_i^{(2)}), z_i),$$

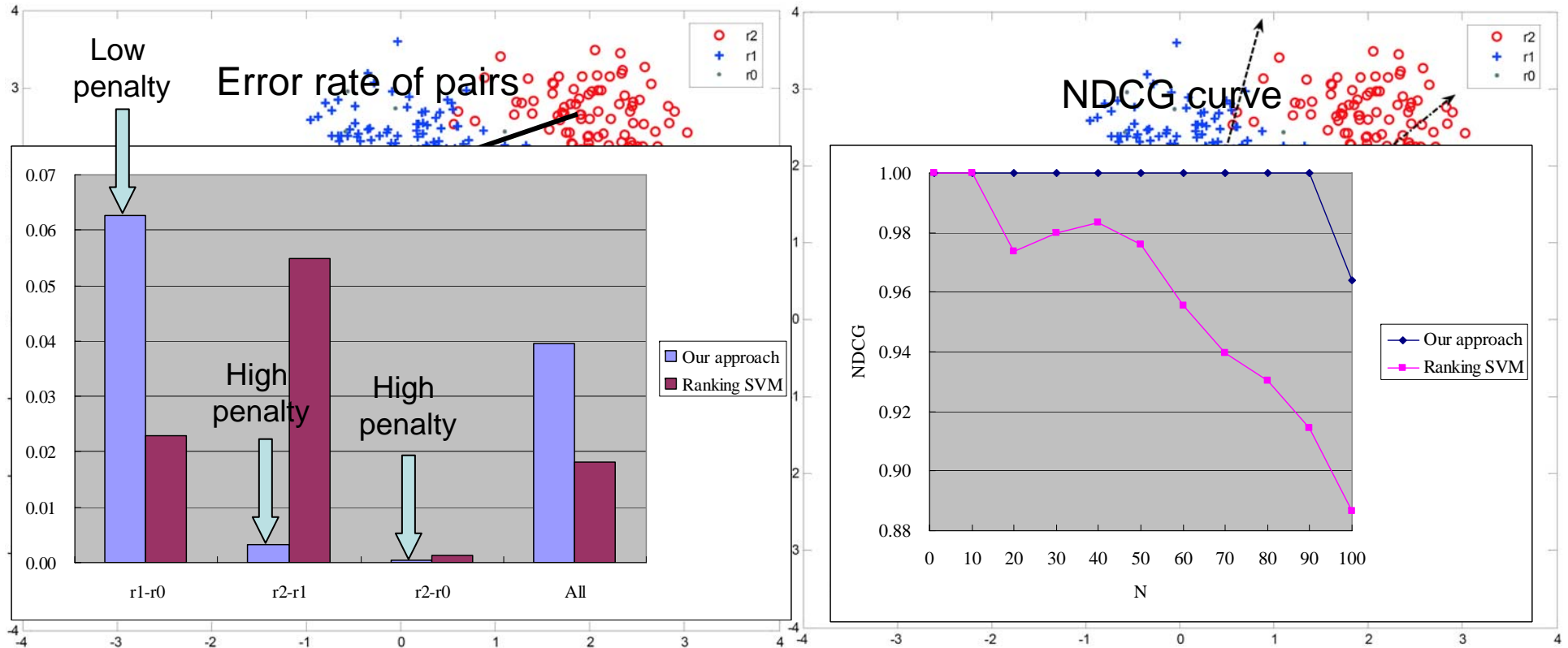
$$\text{where } S' = \left\{ \left((\vec{x}_i, \vec{x}_j), z = \begin{cases} +1 & y_i \succ y_j \\ -1 & y_i \prec y_j \end{cases} \right) : (\vec{x}_i, y_i), (\vec{x}_j, y_j) \in S \right\}$$

Ranking Model

- Model: instances are ranked by $U(x)$



Simulation Experiment 1



Information Desk

(Li et al., 2005)

Information Desk - Microsoft Internet Explorer

File Edit View Favorites Tools Help


Back Forward Stop Home Search Favorites Home Search Go Links msn

Address http://searchlabs1/ Home | About | Feedback

search labs information desk Microsoft

Search:

Who is What is Where is homepage of Who knows about

 Could you please take five minutes to fill in the survey form? [The form>>>](#)

What is...
Definition of a technical term, group name, product name or code name. Expansion of an acronym.

- What is Longhorn?
- What is ATM?

Who is...
Alias, title, department, tel and associated documents of a person.

- Who is Bill Gates?
- Who is steveb?

Where is the homepage of...
Homepage of a group or product.

- Where is the homepage of MSDN?
- Where is the homepage of Office?

Who knows about...
People who know about technical terms, group names or product names.

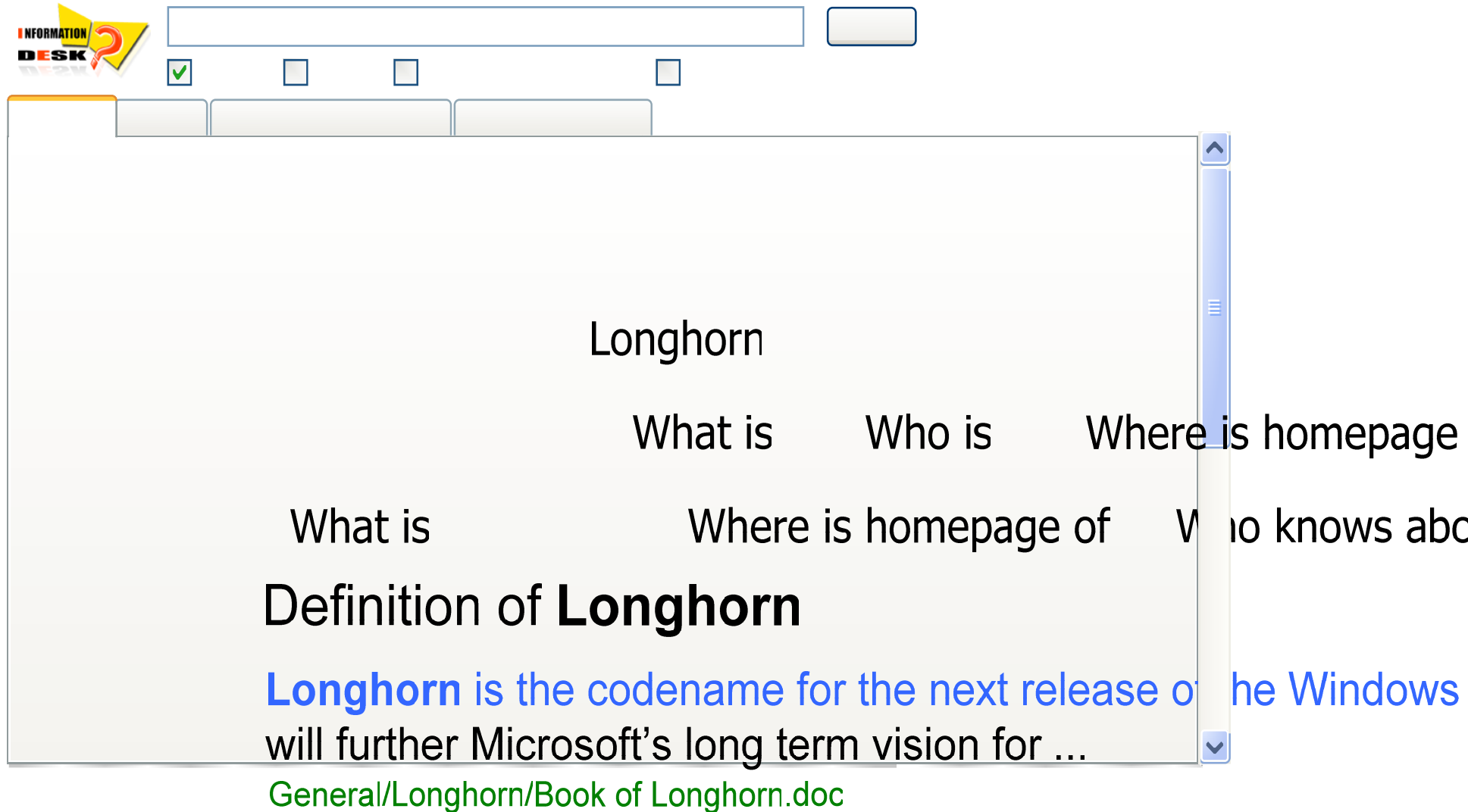
- Who knows about database?
- Who knows about Trustworthy Computing?

Looking for InfoDesk maintained by the sales group? [Find it here](#)

Microsoft Confidential

Done Local intranet

Features -- 'what is'



The screenshot shows a search interface with a search bar at the top containing the text "Longhorn". Below the search bar are several tabs, with the first one selected. The main content area displays search results for "Longhorn". The first result is titled "Definition of **Longhorn**" and includes a blue link: "Longhorn is the codename for the next release of the Windows will further Microsoft's long term vision for ...". Below this link is a green link: "General/Longhorn/Book of Longhorn.doc".

Information Desk

Longhorn

What is Who is Where is homepage

What is Where is homepage of Who knows about


Definition of Longhorn

[Longhorn is the codename for the next release of the Windows will further Microsoft's long term vision for ...](#)

[General/Longhorn/Book of Longhorn.doc](#)

[Longhorn is a platform that enables incredible user experiences](#)

Features – ‘who is’

 Bill Gates

What is Who is Where is homepage of Who knows about

What is | Who is | Where is homepage of | Who knows about

Bill Gates CHRMN & CHIEF SFTWR ARCHITECT


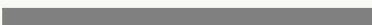

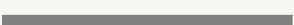
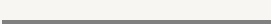
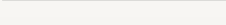
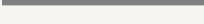
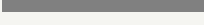
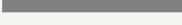
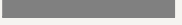
US-Executve-Chairman
+1 (425) XXXXXXXX XXXXXX

Documents of Bill Gates(118)

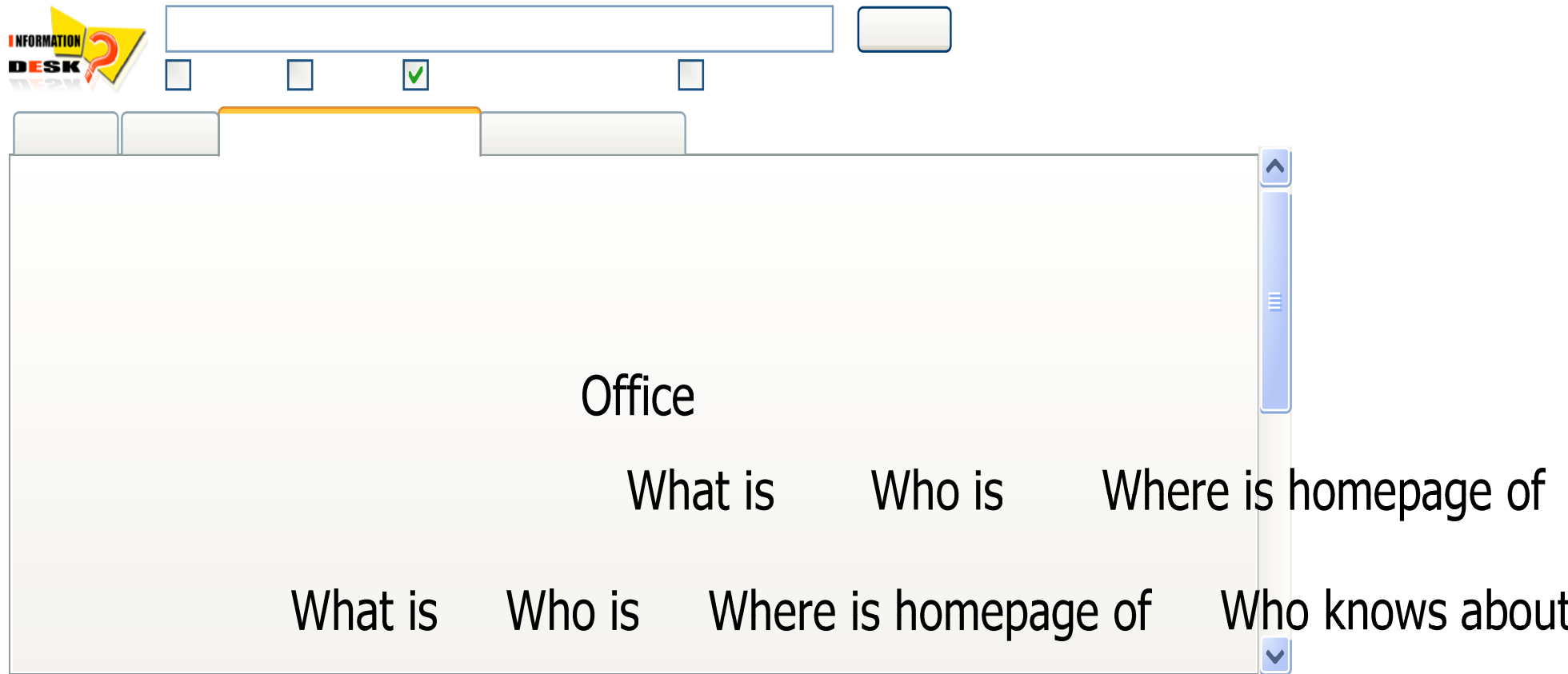
- [My advice to students: Education counts](#)
<http://msmuseum/docs/Myadvicetostudents.doc>
- [Evento NET Reviewers – Seattle – 7/8 Novembro](#)
http://brasil/net/f_docs/NET_Reviewers_Event Anotacoes Dago.doc
- [A Vision for Life Long Learning – Year 2020](#)
http://eduweb/Document Library/Vision for Education_Microsoft_Final.doc
- [Bill Gates answers most frequently asked questions.](#)
<http://msmuseum/docs/BillGatesFAQ.doc>

[>>more](#)

Top 10 terms appearing in documents of Bill Gates

Term 1 (984.4443)	
Term 2 (816.4247)	
Term 3 (595.0771)	
Term 4 (578.5604)	
Term 5 (565.7299)	
Term 6 (435.5366)	
Term 7 (412.4467)	
Term 8 (385.446)	
Term 9 (346.5993)	
Term 10 (345.3285)	

Features – ‘where is homepage of ’



Homepages of **Office**

Office Portal Site

This is the internal site for Office

Features - 'who knows about'

The screenshot shows a web interface for an 'INFORMATION DESK'. At the top left is a logo with a yellow question mark. To its right is a search bar and a button. Below the search bar are four checkboxes, the fourth of which is checked. A horizontal bar below the checkboxes contains several tabs, with the one labeled 'Data Mining' highlighted in orange. The main content area displays search results for 'Data Mining'. The first result is a table with columns: 'What is', 'Who is', and 'Where is homepage of'. The second result is a table with columns: 'What is', 'Who is', 'Where is homepage of', and 'Who knows about'. The third result is a blue link: 'People Associated with Data mining'. The fourth result is a blue link: 'Jamie MacLennan'. Below the search results, the text 'US-SQL Data Warehouse' is displayed, followed by a phone number '+1 (425) XXXXXXXX XXXXXX' and the text 'Associated documents(4):'. On the right side of the interface, there is a vertical scrollbar and a 'DEV' label.

INFORMATION
DESK

What is Who is Where is homepage of Who knows about

People Associated with Data mining

Jamie MacLennan

US-SQL Data Warehouse
+1 (425) XXXXXXXX XXXXXX
Associated documents(4):

DEV

Summary

Summary

- Information retrieval = helping people access information
- Currently search = ranking
- Matching between query and document
- Our work
 - Search by Document Type, Factoid Search
 - New learning algorithm for ranking
- Many issues to study

References

- Brin S. and Page T.(1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine. In : Proceedings of the seventh international conference on World Wide Web.
- Hang Li, Yunbo Cao, Jun Xu, Yunhua Hu, Shenjie Li, and Dmitriy Meyerzon, A New Approach to Intranet Search Based on Information Extraction. Proc. of ACM-CIKM'05 industry track
- Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large Margin Rank Boundaries for Ordinal Regression. . Advances in Large Margin Classifiers (pp. 115-132).
- Joachims T. (2002), Optimizing Search Engines Using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining.
- Lafferty J. and Zhai C. (2001). Document Language Models, Query Models, and Risk Minimization for Information Retrieval. SIGIR
- Ponte J. M. and Croft W. B. (1998). A language modeling approach to information retrieval. In Proceedings of ACM-SIGIR, pp. 275-281.
- Robertson S. E. and Walker S. (1994) Okapi at TREC 3. In Proceedings of TREC
- Yunbo Cao, Jingjing Liu, Shenghua Bao, and Hang Li (2005), Microsoft Research Asia (MSRA) at Enterprise Track of TREC 2005: Expert Search. In Proceedings of TREC.