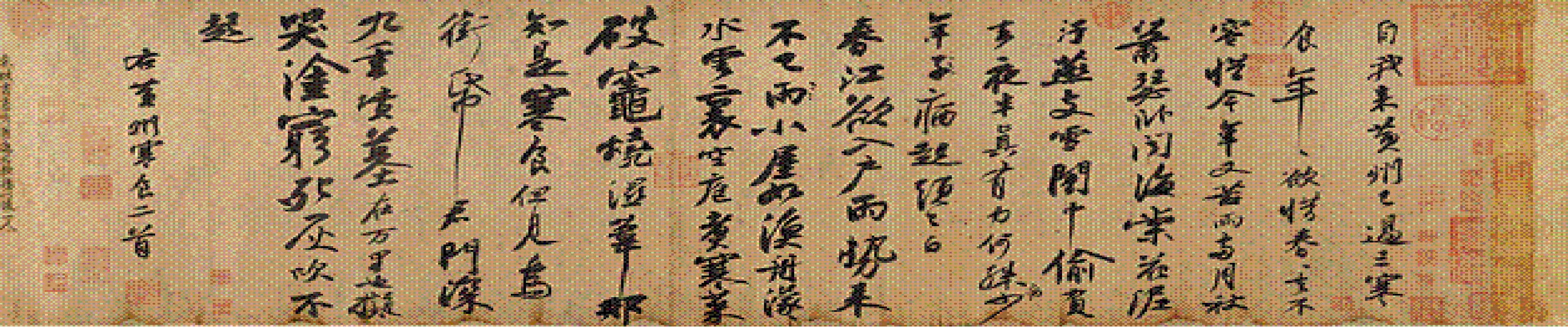


信息检索
Information Retrieval

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA：胡锦毅

Email:hu-jy21@mails.tsinghua.edu.cn

郑重声明

- 此课件仅供选修清华大学计算机系本科生课《信息检索》（课号：40240372）的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

第六章 其它信息检索模型

6.1. Extended Boolean Model

- Salton, Fox and Wu (1983)

$$w_{x,j} = f_{x,j} \times \frac{idf_x}{max_i \ idf_i}$$

$$f_{x,j} = \frac{freq_{x,j}}{max_i \ freq_{i,j}}$$

$$q_{or} = k_x \vee k_y \qquad q_{and} = k_x \wedge k_y$$

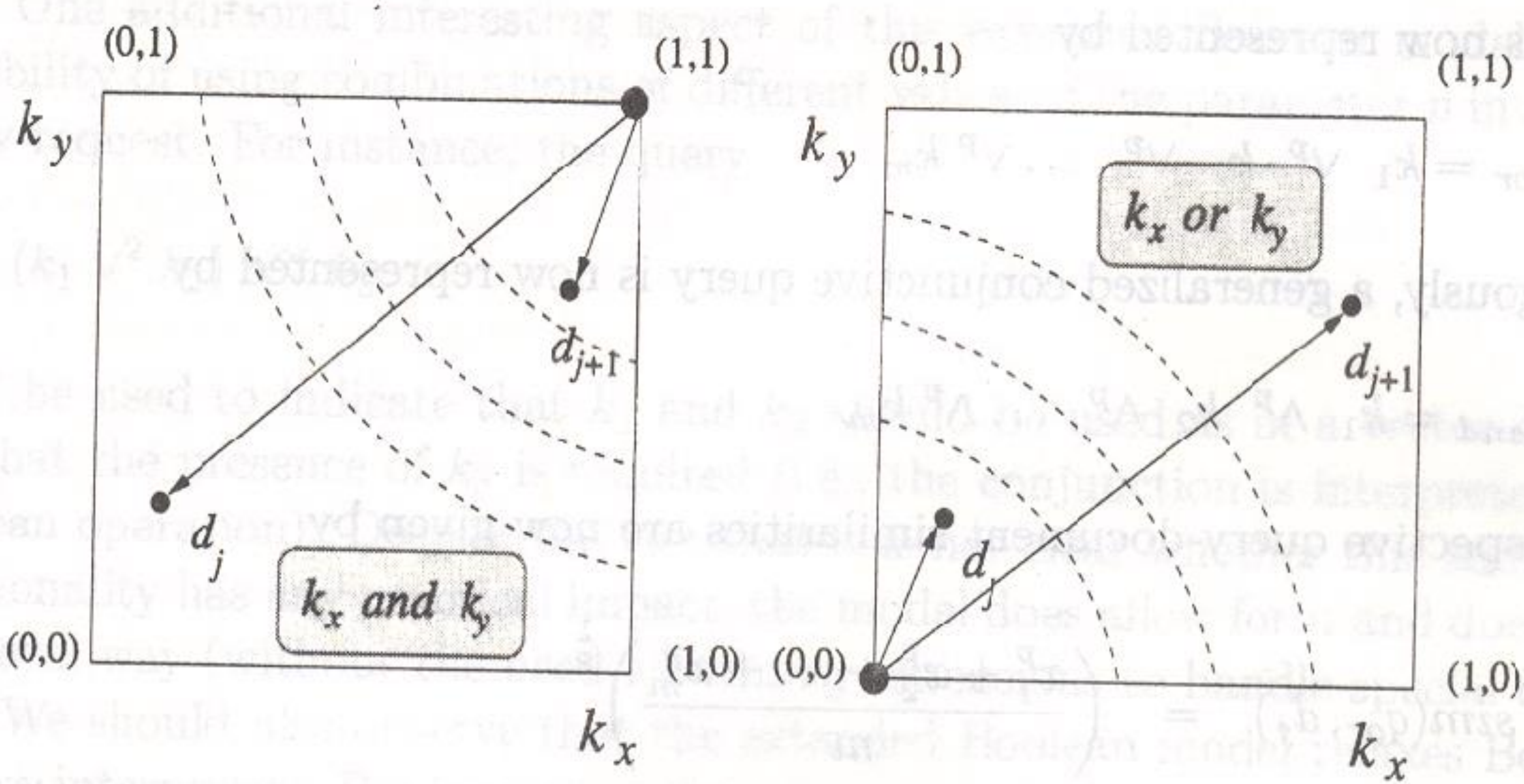
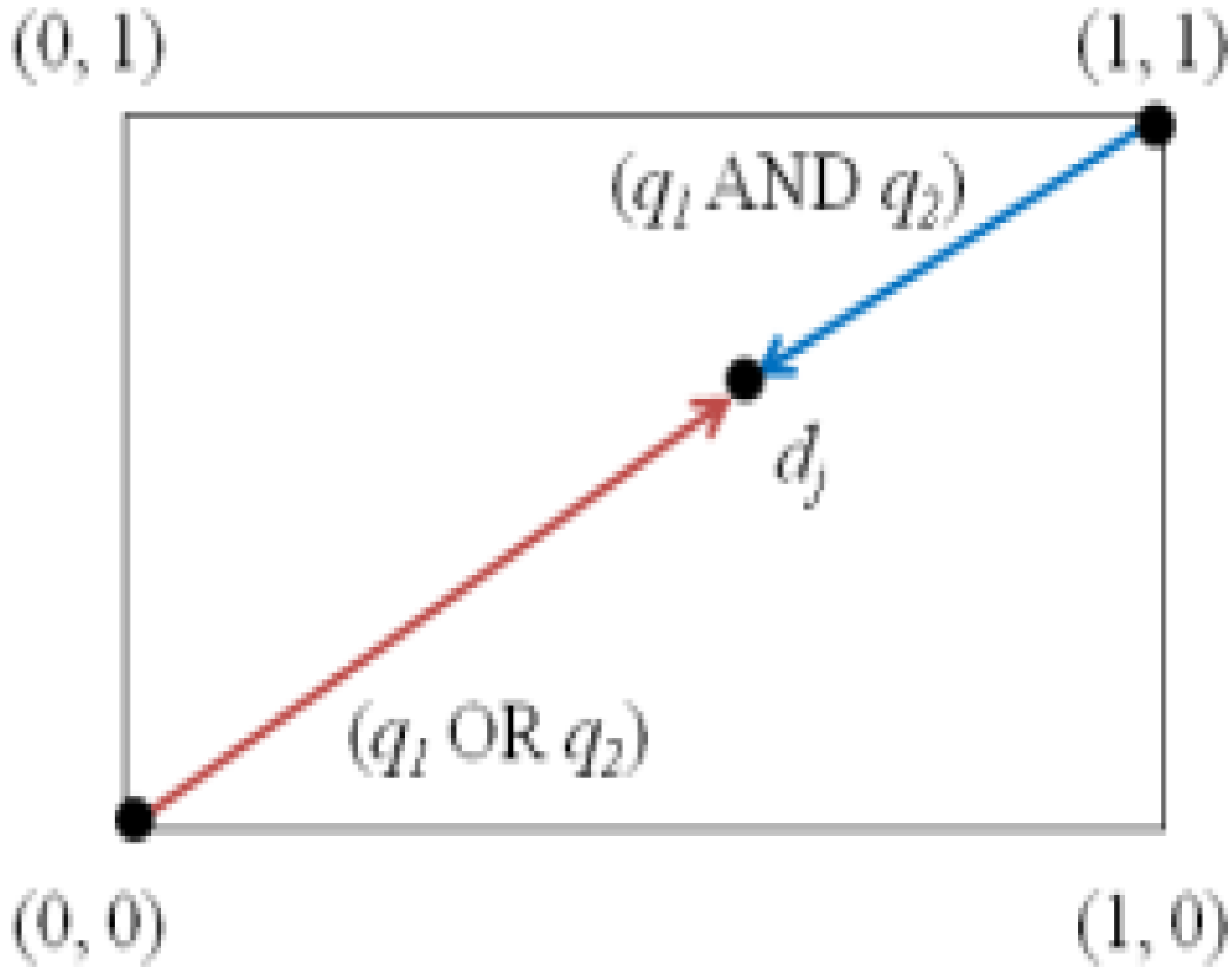


Figure 2.6 Extended Boolean logic considering the space composed of two terms k_x and k_y only.

$$sim(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

$$sim(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$



Term space representation of AND and OR two-term queries.

6.1. Extended Boolean Model

p-norm model

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m \qquad q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$$

$$sim(q_{or}, d_j) = \left(\frac{x_1^p + x_2^p + \ldots + x_m^p}{m} \right)^{\frac{1}{p}}$$

$$sim(q_{and}, d_j) = 1 - \left(\frac{(1-x_1)^p + (1-x_2)^p + \ldots + (1-x_m)^p}{m} \right)^{\frac{1}{p}}$$

p=1

$$sim(q_{or}, d_j) = sim(q_{and}, d_j) = \frac{x_1 + \ldots + x_m}{m}$$

P=无穷

$$sim(q_{or}, d_j) = max(x_i)$$
$$sim(q_{and}, d_j) = min(x_i)$$

6.1. Extended Boolean Model

$$q = (k_1 \wedge^p k_2) \vee^p k_3$$

$$sim(q,d) = \left(\frac{\left(1 - \left(\frac{(1-x_1)^p + (1-x_2)^p}{2}\right)^{\frac{1}{p}}\right)^p + x_3^p}{2} \right)^{\frac{1}{p}}$$

$$(k_1 \vee^2 k_2) \wedge^\infty k_3$$

6.2. Probabilistic Model

- Roberston and Sparck Jones (1976)

  The *binary independence retrieval* model

Assumption (Probabilistic Principle) Given a user query q and a document d_j in the collection, the probabilistic model tries to estimate the probability that the user will find the document d_j interesting (i.e., relevant). The model assumes that this probability of relevance depends on the query and the document representations only. Further, the model assumes that there is a subset of all documents which the user prefers as the answer set for the query q. Such an *ideal* answer set is labeled R and should maximize the overall probability of relevance to the user. Documents in the set R are predicted to be *relevant* to the query. Documents not in this set are predicted to be *non-relevant*.

Definition *For the probabilistic model, the index term weight variables are all binary i.e., $w_{i,j} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$. A query q is a subset of index terms. Let R be the set of documents known (or initially guessed) to be relevant. Let $\overline{R}$ be the complement of R (i.e., the set of non-relevant documents). Let $P(R|\vec{d}_j)$ be the probability that the document d_j is relevant to the query q and $P(\overline{R}|\vec{d}_j)$ be the probability that d_j is non-relevant to q. The similarity $sim(d_j, q)$ of the document d_j to the query q is defined as the ratio*

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\overline{R}|\vec{d}_j)}$$

Using Bayes' rule,

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) \times P(R)}{P(\vec{d}_j|\overline{R}) \times P(\overline{R})}$$

$P(\vec{d}_j|R)$ stands for the probability of randomly selecting the document d_j from the set R of relevant documents. Further, $P(R)$ stands for the probability that a document randomly selected from the entire collection is relevant. The meanings attached to $P(\vec{d}_j|\overline{R})$ and $P(\overline{R})$ are analogous and complementary.

Since $P(R)$ and $P(\overline{R})$ are the same for all the documents in the collection, we write,

$$sim(d_j, q) \sim \frac{P(\vec{d_j}|R)}{P(\vec{d_j}|\overline{R})}$$

Assuming independence of index terms,

$$sim(d_j, q) \sim \frac{(\prod_{g_i(\vec{d_j})=1} P(k_i|R)) \times (\prod_{g_i(\vec{d_j})=0} P(\overline{k}_i|R))}{(\prod_{g_i(\vec{d_j})=1} P(k_i|\overline{R})) \times (\prod_{g_i(\vec{d_j})=0} P(\overline{k}_i|\overline{R}))}$$

$P(k_i|R)$ stands for the probability that the index term k_i is present in a document randomly selected from the set R. $P(\overline{k}_i|R)$ stands for the probability that the index term k_i is not present in a document randomly selected from the set R. The probabilities associated with the set $\overline{R}$ have meanings which are analogous to the ones just described.

Taking logarithms, recalling that $P(k_i|R) + P(\bar{k}_i|R) = 1$, and ignoring factors which are constant for all documents in the context of the same query, we can finally write

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\overline{R})}{P(k_i|\overline{R})} \right)$$

初始：

$$P(k_i|R) = 0.5$$
$$P(k_i|\overline{R}) = \frac{n_i}{N}$$

更新：

Let V be a subset of the documents initially retrieved and ranked by the probabilistic model. Such a subset can be defined, for instance, as the top r ranked documents where r is a previously defined threshold. Further, let V_i be the subset of V composed of the documents in V which contain the index term k_i. For simp … o the number of elements in

$$P(k_i|R) = \frac{V_i}{V}$$
$$P(k_i|\overline{R}) = \frac{n_i - V_i}{N - V}$$

6.2. Probabilistic Model

平滑（例如：当V_i=0）

$$P(k_i|R) = \frac{V_i + 0.5}{V + 1}$$

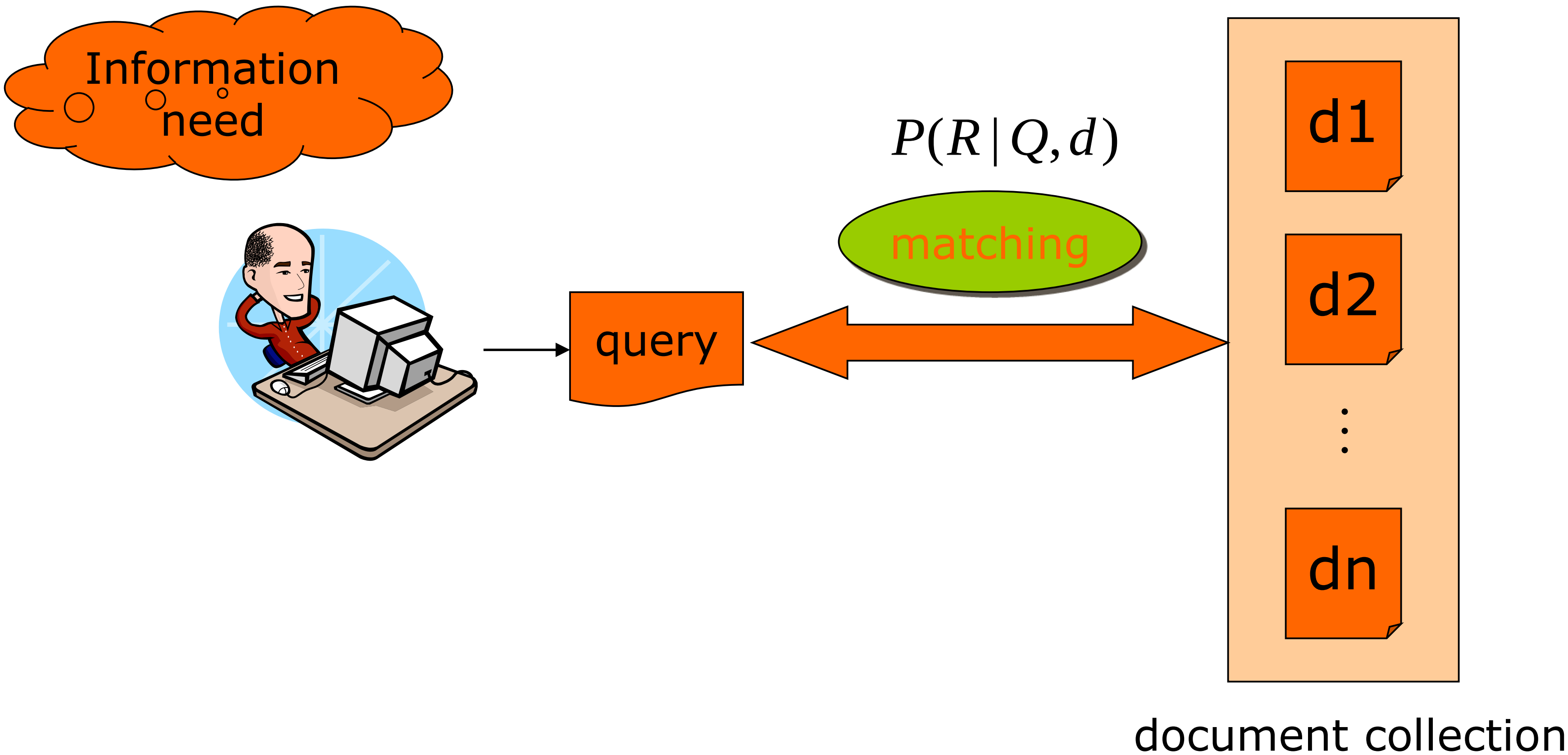
$$P(k_i|\overline{R}) = \frac{n_i - V_i + 0.5}{N - V + 1}$$

$$P(k_i|R) = \frac{V_i + \frac{n_i}{N}}{V + 1}$$

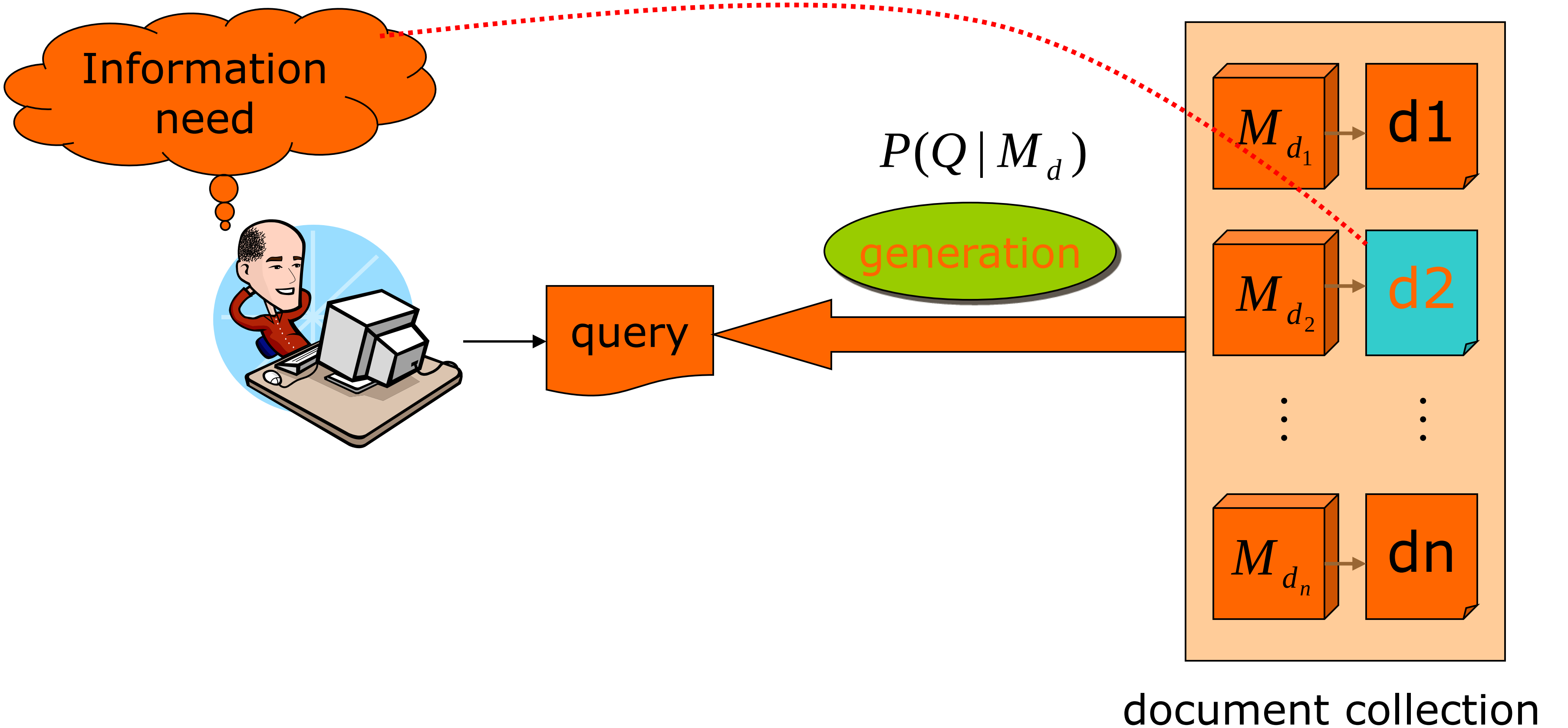
$$P(k_i|\overline{R}) = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1}$$

6.3. Language Model

Standard Probabilistic IR



6.3. Language Model



6.3. Language Model

Models *probability* of generating strings in the language (commonly all strings over Σ)

Model M

0.2	the
0.1	a
0.01	man
0.01	woman
0.03	said
0.02	likes
…	

the	man	likes	the	woman
0.2	0.01	0.02	0.2	0.01

multiply

P(s | M) = 0.00000008

6.3. Language Model

Model *probability* of generating any string

Model M1

0.2	the
0.01	class
0.0001	sayst
0.0001	pleaseth
0.0001	yon
0.0005	maiden
0.01	woman

Model M2

0.2	the
0.0001	class
0.03	sayst
0.02	pleaseth
0.1	yon
0.01	maiden
0.0001	woman

the	class	pleaseth	yon	maiden
0.2	0.01	0.0001	0.0001	0.0005
0.2	0.0001	0.02	0.1	0.01

$P(s|M2) > P(s|M1)$

6.3. Language Model

- For any sentence $S = w_1, \ldots\ldots, w_t$

$$PROB(S) = PROB(w_1, \ldots\ldots, w_t)$$

$$= PROB(w_1, \ldots\ldots, w_{t-1}) \times PROB(w_t \mid w_1, \ldots\ldots, w_{t-1})$$
$$= PROB(w_1, \ldots\ldots, w_{t-2}) \times PROB(w_{t-1} \mid w_1, \ldots\ldots, w_{t-2}) \times PROB(w_t \mid w_1, \ldots\ldots, w_{t-1})$$
$$= \ldots\ldots$$
$$= PROB(w_1, w_2) \times PROB(w_3 \mid w_1, w_2) \times \ldots\ldots \times PROB(w_{t-1} \mid w_1, \ldots\ldots, w_{t-2}) \times PROB(w_t \mid w_1, \ldots\ldots, w_{t-1})$$
$$= PROB(w_1) \times PROB(w_2 \mid w_1) \times PROB(w_3 \mid w_1, w_2) \times \ldots\ldots \times PROB(w_{t-1} \mid w_1, \ldots\ldots, w_{t-2}) \times PROB(w_t \mid w_1, \ldots\ldots, w_{t-1})$$

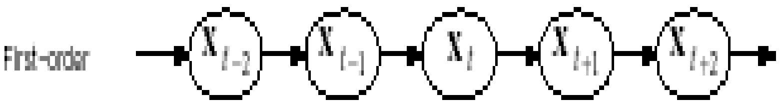
6.3. Language Model

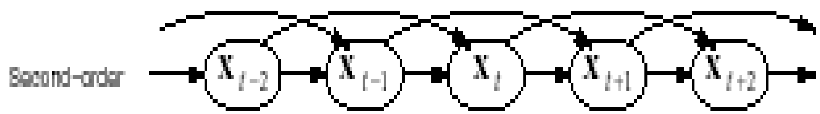
n-gram models:

unigram(The 0 order Markov model): $PROB(w_i)$

bigram(The first oder Markov Model): $PROB(w_i \mid w_{i-1})$

trigram(The second order Markov Model): $PROB(w_i \mid w_{i-1}, w_{i-2})$





Using unigram approximation:

$$PROB(w_1, \ldots\ldots, w_t) \cong PROB(w_1) \times PROB(w_2) \times \ldots\ldots \times PROB(w_t)$$

$$= \prod_{i=1,t} PROB(w_i)$$

Using bigram approximation:

$$PROB(w_1, \ldots\ldots, w_t) \cong PROB(w_1) \times PROB(w_2 \mid w_1) \times \ldots\ldots \times PROB(w_t \mid w_{t-1})$$

$$= PORB(w_1) \prod_{i=2,t} PROB(w_i \mid w_{i-1})$$

6.3. Language Model

* Treat each document as the basis for a model (e.g., unigram sufficient statistics)
* Rank document d based on P(d | q)

  P(d | q) = P(q | d) x P(d) / P(q)

  -- P(q) is the same for all documents, so ignore

  -- P(d) [the prior] is often treated as the same for all d

  -- P(q | d) is the probability of q given d's model

6.3. Language Model

* Language Modeling Approach
  -- Attempt to **model query generation process**
  -- Documents are ranked by **the probability that a query would be observed as a random sample from the respective document model**

$$P(Q|M_D) = \prod_{w \in Q} P(w|M_D) \prod_{w \notin Q} (1 - P(w|M_D))$$

  -- Usually a unigram estimate of words is used
      Some work on bigrams

6.3. Language Model

* The probability of producing the query given the language model of document d using MLE is:

$$\hat{p}(Q \mid M_d) = \prod_{t \in Q} \hat{p}_{ml}(t \mid M_d)$$

$$= \prod_{t \in Q} \frac{tf_{(t,d)}}{dl_d}$$

Unigram assumption:
Given a particular language model, the query terms occur independently

M_d : language model of document d

$tf_{(t,d)}$: raw tf of term t in document d

dl_d : total number of tokens in document d

6.3. Language Model

* Zero probability $p(t \mid M_d) = 0$

  -- May not wish to assign a probability of zero to a document that is missing one or more of the query terms

* General pproach

  -- A non-occurring term is possible, but no more likely than would be expected by chance in the collection. $tf_{(t,d)} = 0$ $p(t \mid M_d) = \frac{cf_t}{cs}$

  -- If

  cf_t : raw count of term t in the collection

  cs : raw collection size(total number of tokens in the collection)

6.3. Language Model

Mixture model

$$p(Q \mid d) = \prod_{t \in Q} ((1-\lambda)\, p(t) + \lambda p(t \mid M_d))$$

general language model

individual-document model

* Mixes the probability from the document with the general collection frequency of the word.
* Correctly setting λ is very important
* Can tune λ to optimize performance

6.3. Language Model

Ponte and Croft Experiments

* Data
  - -- TREC topics 202-250 on TREC disks 2 and 3
    - Natural language queries consisting of one sentence each
  - -- TREC topics 51-100 on TREC disk 3 using the concept fields
    - Lists of good terms

<num>Number: 054

<dom>Domain: International Economics

<title>Topic: Satellite Launch Contracts

<desc>Description:
… </desc>

<con>Concept(s):

1. Contract, agreement
2. Launch vehicle, rocket, payload, satellite
3. Launch services, … </con>

6.3. Language Model

Precision/ recall results 202-250

	tf.idf	LM	%chg	I/D	Sign	Wilc.
Rel:	6501	6501				
Rret.:	3201	3364	+5.09	36/43	0.0000⋆	0.0002⋆
Prec.						
0.00	0.7439	0.7590	+2.0	10/22	0.7383	0.5709
0.10	0.4521	0.4910	+8.6	24/42	0.2204	0.0761
0.20	0.3514	0.4045	+15.1	27/44	0.0871	0.0081⋆
0.30	0.2761	0.3342	+21.0	28/43	0.0330⋆	0.0054⋆
0.40	0.2093	0.2572	+22.9	25/39	0.0541	0.0158⋆
0.50	0.1558	0.2061	+32.3	24/35	0.0205⋆	0.0018⋆
0.60	0.1024	0.1405	+37.1	22/27	0.0008⋆	0.0027⋆
0.70	0.0451	0.0760	+68.7	13/15	0.0037⋆	0.0062⋆
0.80	0.0160	0.0432	+169.6	9/10	0.0107⋆	0.0035⋆
0.90	0.0033	0.0063	+89.3	2/3	0.5000	undef
1.00	0.0028	0.0050	+76.9	2/3	0.5000	undef
Avg:	0.1868	0.2233	+19.55	32/49	0.0222⋆	0.0003⋆
Prec.						
5	0.4939	0.5020	+1.7	10/21	0.6682	0.4106
10	0.4449	0.4898	+10.1	22/30	0.0081⋆	0.0154⋆
15	0.3932	0.4435	+12.8	19/26	0.0145⋆	0.0038⋆
20	0.3643	0.4051	+11.2	22/34	0.0607	0.0218⋆
30	0.3313	0.3707	+11.9	28/41	0.0138⋆	0.0070⋆
100	0.2157	0.2500	+15.9	32/42	0.0005⋆	0.0003⋆
200	0.1655	0.1903	+15.0	35/44	0.0001⋆	0.0000⋆
500	0.1004	0.1119	+11.4	36/44	0.0000⋆	0.0000⋆
1000	0.0653	0.0687	+5.1	36/43	0.0000⋆	0.0002⋆
RPr	0.2473	0.2876	+16.32	34/43	0.0001⋆	0.0000⋆

6.3. Language Model

Precision/ recall results 51-100

	tf.idf	LM	%chg	I/D	Sign	Wilc.
Rel:	10485	10485				
Rret.:	5818	6105	+4.93	32/42	0.0005⋆	0.0003⋆
Prec.						
0.00	0.7274	0.7805	+7.3	10/22	0.7383	0.2961
0.10	0.4861	0.5002	+2.9	26/44	0.1456	0.1017
0.20	0.3898	0.4088	+4.9	24/45	0.3830	0.1405
0.30	0.3352	0.3626	+8.2	28/47	0.1215	0.0277⋆
0.40	0.2826	0.3064	+8.4	25/45	0.2757	0.0286⋆
0.50	0.2163	0.2512	+16.2	26/40	0.0403⋆	0.0007⋆
0.60	0.1561	0.1798	+15.2	20/30	0.0494⋆	0.0025⋆
0.70	0.0913	0.1109	+21.5	14/22	0.1431	0.0288⋆
0.80	0.0510	0.0529	+3.7	8/13	0.2905	0.2108
0.90	0.0179	0.0152	−14.9	1/4	0.3125	undef
1.00	0.0005	0.0004	−11.9	1/2	0.7500	undef
Avg:	0.2286	0.2486	+8.74	32/50	0.0325⋆	0.0015⋆
Prec.						
5	0.5320	0.5960	+12.0	15/21	0.0392⋆	0.0125⋆
10	0.5080	0.5260	+3.5	14/30	0.7077	0.1938
15	0.4933	0.5053	+2.4	14/28	0.5747	0.3002
20	0.4670	0.4890	+4.7	16/34	0.6962	0.1260
30	0.4293	0.4593	+7.0	20/32	0.1077	0.0095⋆
100	0.3344	0.3562	+6.5	29/45	0.0362⋆	0.0076⋆
200	0.2670	0.2852	+6.8	29/44	0.0244⋆	0.0009⋆
500	0.1797	0.1881	+4.7	30/42	0.0040⋆	0.0011⋆
1000	0.1164	0.1221	+4.9	32/42	0.0005⋆	0.0003⋆
RPr	0.2836	0.3013	+6.24	30/43	0.0069⋆	0.0052⋆