Presentation for 2nd Discipline Construction Conference

Introduction of IR and Our Works

Hao-ming WANG

Dept. of Computer Science
Xi’an University of Finance & Economics

hmwang@mail.xaufe.edu.cn

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 1 of 31
Go Back
Full Screen
Close
Quit

Contents

- Part 1 Introduction
- Part 2 Algorithm of the CLBCRA
- Part 3 Experimental for CLBCRA
- Part 4 Conclusion and Future work
- Part 5 Appendix

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 2 of 31
Go Back
Full Screen
Close
Quit

1 Introduction

Information retrieval (IR) is the science of

- Searching for information in documents ;
- Searching for documents themselves;
- Searching for metadata which describe documents;
- Searching within databases.

whether in

- Relational stand-alone databases;
- hypertextually-networked databases.

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 3 of 31
Go Back
Full Screen
Close
Quit

IR is **interdisciplinary**, based on

- Computer Science;
- Mathematics;
- Library Science;
- Information Science;
- Information Architecture;
- Cognitive Psychology(**认知心理学**);
- Linguistics(**语言学**);
- Statistics;
- Physics

In our project, we discuss the IR in **Internet**.

There are many search engines, such as Yahoo, Google, etc. to help the user to search and collect the information from the Internet.

There are 2 kinds of Search Engines, based on

- Content: such as Yahoo;
- Link: such as Google;

The features of the Internet,

- mass;
- semi-structure;

have become drawbacks in using the information widely in Internet.

Our work is constructing the new model of combining the content and the link of the pages in order to compute the **value** of pages.

1.1. Precision and Recall

- 查全率(Recall):它反映该系统文献库中实有的相关文献量在多大程度上被检索出来。

$$\text{查全率} = \frac{\text{检出相关文献量}}{\text{文献库内相关文献总量}} \times 100\%.$$

$$Recall = \frac{Ret \cap Rele}{Rele} \times 100\%.$$

- 查准率(Precision):它反映每次从该系统文献库中实际检出的全部文献中有多少是相关的。

$$\text{查准率} = \frac{\text{检出相关文献量}}{\text{检出文献总量}} \times 100\%.$$

$$Precision = \frac{Ret \cap Rele}{Ret} \times 100\%.$$

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 6 of 31
Go Back
Full Screen
Close
Quit

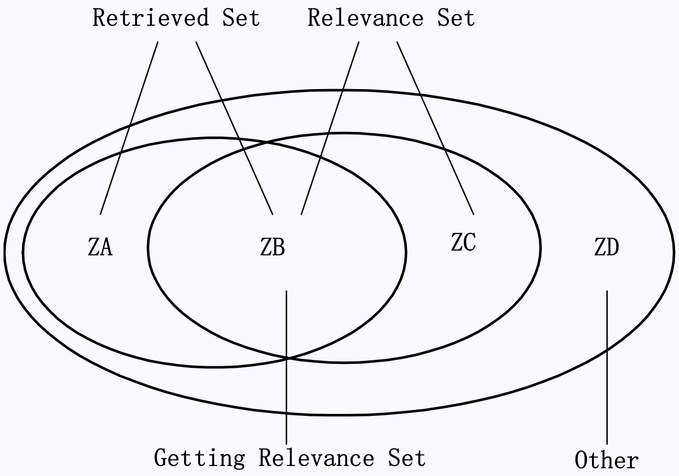


Figure 1: Concept of Information Retrieval

- $ZA \cup ZB$: all retrieval pages set; $ZB \cup ZC$: all relevance pages set;
- ZA: pages set which retrieved but not relevance to the query;
- ZB: pages set which retrieved and relevance to the query indeed;
- ZC: pages set which relevance but could be retrieved;
- ZD: all the other pages set;
- Precision: $\dfrac{ZB}{ZA+ZB}$; Recall: $\dfrac{ZB}{ZB+ZC}$.

1.2. TFIDF

TFIDF (Term Frequency / Inverse Document Frequency) is the most common weighting method used to describe documents in the Vector Space Model (VSM), particularly in IR problems.

Assuming vector $\tilde{d} = (d^{(1)}, d^{(2)}, ..., d^{(n)})$ represents the document d in a vector space. Where $d^{(i)}(i \in (0, n))$ is the weight of the term w_i appeared in document d. d_i is calculated as a combination of the statistics $TF(w, d)$ and $DF(w)$ (document frequency).

$$IDF(w) = log\frac{N_{all}}{DF(w)}. \quad d^{(i)} = TF(w_i, d) \times IDF(w_i).$$

$$Similarity(d', C) = cos(d', C) = argmax(\frac{\tilde{d}' \cdot \tilde{C}}{\|\tilde{d}'\| \cdot \|\tilde{C}\|})$$

$$= argmax(\frac{\sum_{i=1}^{|F|}[d'^{(i)} \cdot C^{(i)}]}{\sqrt{\sum_{i=1}^{|F|}[d'^{(i)}]^2} \cdot \sqrt{\sum_{i=1}^{|F|}[C^{(i)}]^2}}).$$

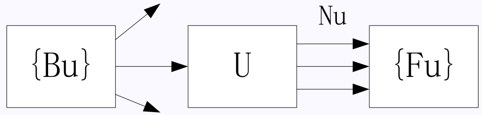
1.3. Pagerank

- **Define**

  PageRank algorithm first introduced by Brin and Page:

  Let u be the web page. Then let F_u be the set of pages u points to and B_u be the set of pages that point to u. Let N_u be the number of links from u and let c be a factor used for normalization (so that the total rank of all web pages is constant):

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}.$$



- **Link Matrix**

  Considering the pages and the links as a graph $G = P(Page, Link)$, defined

$$p_{ij} = \begin{cases} 1 & \exists(Link\, i \rightarrow j\,) \\ 0 & Otherwise. \end{cases}$$

  P corresponds to a **Markov chain**.

- **Transfer probability Matrix**

$$p_{ij} = \begin{cases} 1/deg(i) & \exists(Link\, i \rightarrow j\,) \\ 0 & Otherwise. \end{cases}$$

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 9 of 31
Go Back
Full Screen
Close
Quit

- **Dangling Page**

  Pages has not any out-link, which means *deg(i)=0*. The matrix P is not a row-stochastic.

  Changing P to $P' = P + d \cdot v^T$, Where

$$d = \begin{cases} 1 & if \quad deg(i) = 0 \\ 0 & Otherwise. \end{cases}$$

  is the dangling page indictor.

- **Chapman-Kolmogorov Equations**

  For the Markov chains,

$$P^{(n+m)} = P^{(n)} \cdot P^{(m)};$$

$$P^{(2)} = P^{(1)} \cdot P^{(1)} = P \cdot P = P^2;$$

$$P^{(n)} = P^{(n-1+1)} = P^{(n-1)} \cdot P^{(1)} = P^{n-1} \cdot P = P^n.$$

  That means, the $n - step$ transition matrix can be obtained by multiplying the matrix P by itself n times.

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 10 of 31
Go Back
Full Screen
Close
Quit

- **Irreducible(不可约)**

  The irreducible property can be described as: we have a Markov chain, if it could not be divided into $n(n>1)$ parts. We call the Markov chain has the irreducible property.

- **Aperiodic(非周期)**

  When the matrix P is irreducible, the highest common divisor τ of all position integers k such that $P^k(i,i)>0$ for all $i=(1,q)$ is called the period of P. The matrix P is called aperiodic if $\tau=1$

- **Changing Transfer probability Matrix**

  As the existing of zero entries in the matrix P'. P' can be modified by adding the connection between every pair of pages.

$$Q=P''=cP'+(1-c)ev^T, \quad e=(1,1,\cdots,1)^T.$$

  Where c is called dangling factor, and $c\in(0,1)$. In most of the references, the c is set [0.85,1).

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 11 of 31
Go Back
Full Screen
Close
Quit

- **Converge**

  For matrix Q: (Primitive matrix:本原矩阵)

  (1) Irreducible: Strong connect ;

  (2) Aperiodic: $Q_{ii}^{(k)} > 0, (i, k \in [1, n])$;

  (3) Row-stochastic ;

  The Perron-Frobenius theorem guarantees the equation

  $$x^{(k+1)} = Q^T x^{(k)}$$

  (for the eigensystem $Q^T x = x$) converges to the principal eigenvector with eigenvalue 1, and there is a real, positive, and the biggest eigenvalue.

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 12 of 31
Go Back
Full Screen
Close
Quit

- **Influence of Changing the Link Matrix**

  Changing the link (transition) matrix P to Q in 2 steps:

  - guarantees the matrix is row-stochastic by divided by the out-link number of each page;
  - guarantees the matrix is irreducible by adding the link pair to each page.

  The second step adds the link between all pages. Is it possible that the modification of matrix changes the eigenvalue order of the matrix or changes the importance of the pages?

  In Ref.(Faults of PageRank / Something is Wrong with Google’ s Mathematical Model), the author points out a example. And then the author goes on to explain a new algorithm with the same complexity of the original PageRank algorithm that solves this problem.

2 Content and Link Based Complete Ranking Algorithm: CLBCRA

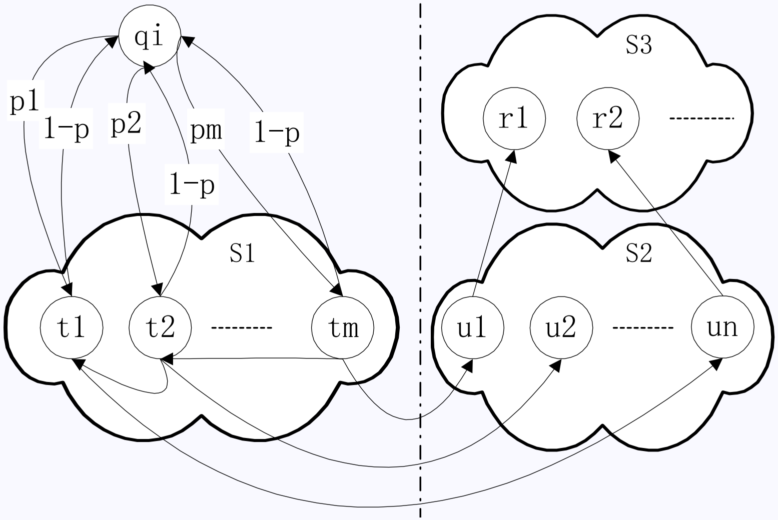


Figure 2: New Model

There are 4 kinds of nodes in the model, they are,

- Set S_0: $q_1, q_2, \cdots, q_1 00$;
- Set S_1: the pages can be reached in 1 step from the query; just as the $d_{11}, d_{12}, \cdots, d_{1m}$;
- Set S_2: the pages can be reached in 1 step from the pages in set S_1, just as the $d_{21}, d_{22}, \cdots, d_{2n}$;
- Set S_3: all the other pages which are not belonged to S_1 and S_2, just as $d_{31}, d_{32}, \cdots$;

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Go Back
Full Screen
Close
Quit

2.1. $\rho 1$ and $\rho 2$

- $$\rho_1 = \frac{Rele \cap Ret}{Ret} = p_i. \tag{1}$$

- $$\rho_2 = \frac{Rele \cap \overline{Ret}}{\overline{Ret}} = \frac{Rele - Ret \cap Rele}{N - Ret} = \frac{Rele - Rele * RE}{N - Ret}$$
$$= \frac{Rele * (1 - RE)}{N - Ret} = \frac{(1 - RE) * \frac{PR * Ret}{RE}}{N - Ret} = PR * \frac{1 - RE}{RE} * \frac{\frac{Ret}{N}}{1 - \frac{Ret}{N}}. \tag{2}$$

Ret: the number of total pages which the retrieval system can get for the given query;
$Rele$: the number of pages which relevant to the given query;
N: the number of total pages, which is a very large number;
PR: the precision of retrieval system;
RE: the recall of retrieval system.

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 15 of 31
Go Back
Full Screen
Close
Quit

2.2. Transfer probability Matrix

- $(t_{0i}, \forall i)$: the probability from query q_i to page $i, i \in S_1$.

$$t_{0i} = \frac{\delta_{0i}}{\sum\limits_{i} \delta_{0i}}. \tag{3}$$

  where $\delta_{0i} = Relation(q_i, t_i), (q_i \in S_0) \wedge (t_i \in S_1)$;

- $(t_{ij}, \forall i, j)$: the probability from page $t_i \rightarrow t_j$, t_i is relevant to q_i;

$$t_{ij} = \begin{cases} \rho_1 * m_{ij} & p_i(q) > 0; \\ \rho_2 * m_{ij} & Otherwise. \end{cases} \tag{4}$$

  where $m_{ij} = \dfrac{1}{\sum\limits_{j} linknum(i \rightarrow j)}$.

- $(t_{i0}, \forall i)$: the probability of returning to query when the page t_i is not relevant to the query.

$$t_{i0} = \begin{cases} 1 - \rho_1 & p_i(q) > 0; \\ 1 - \rho_2 & Otherwise. \end{cases} \tag{5}$$

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 16 of 31
Go Back
Full Screen
Close
Quit

Transfer Probability Matrix,

$$\mathbf{T} = \begin{pmatrix} 0 & P'(q) & 0 & 0 \\ (1-\rho_1)U_1 & \rho_1 * M_{11} & \rho_1 * M_{12} & 0 \\ (1-\rho_2)U_2 & \rho_2 * M_{21} & \rho_2 * M_{22} & \rho_2 * M_{23} \\ (1-\rho_2)U_3 & \rho_2 * M_{31} & \rho_2 * M_{32} & \rho_2 * M_{33} \end{pmatrix} \tag{6}$$

when $\rho_2 \ll 1$,

$$\mathbf{T} = \begin{pmatrix} 0 & P'(q) & 0 & 0 \\ (1-\rho_1)U_1 & \rho_1 * M_{11} & \rho_1 * M_{12} & 0 \\ U_2 & 0 & 0 & 0 \\ U_3 & 0 & 0 & 0 \end{pmatrix}.$$

$$\Rightarrow \mathbf{T}' = \begin{pmatrix} 0 & (1-\rho_1)U_1 & U_2 & U_3 \\ P(q) & \rho_1 * M'_{11} & 0 & 0 \\ 0 & \rho_1 * M'_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{7}$$

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 17 of 31
Go Back
Full Screen
Close
Quit

2.3. Computing the Eigenvalue

Assuming $QQ = (x_0, X_1', X_2', X_3')'$, From $T' * QQ = QQ$,

$$\begin{pmatrix} 0 & (1-\rho_1)U_1 & U_2 & U_3 \\ P(q) & \rho_1 * M_{11}' & 0 & 0 \\ 0 & \rho_1 * M_{12}' & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_0 \\ X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} x_0 \\ X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

We get

$$\begin{cases} x_0 = (2 + \rho_1 * |V|)^{-1} \\ X_1 = x_0 * V \\ X_2 = x_0 * \rho_1 * M_{12}' * V \\ X_3 = 0. \end{cases}$$

Where $V = (I - \rho_1 * M_{11}')^{-1} * P(q)$.

We set

$$S = S_1 \cup S_2.$$

3 Experimental for CLBCRA

3.1. WT10g

- 1.6M个页面，每个页面都有一个名字;
- 100 个模拟查询: 编号从451 到550;
- 与模拟查询的相关性信息: 每个查询与每个页面, 0 = 不相关; 1 = 相关;
- 链出信息;
- 链入信息.

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 19 of 31
Go Back
Full Screen
Close
Quit

3.2. Constructing the test Data-Set

(1) Selecting all 100 queries $q_i, (i \in [1, 100])$ orderly;

(2) Computing the $TFIDF$ value of the query q_i to all pages in WT10g; Selecting the $Top - N, (N = 500, 1000, 5000, 10^4, 1.5 * 10^4, 3 * 10^4)$ pages to construct the data set $S_{1i}, (i = [1, 6])$ respectively;

(3) Drawing up all links

$$L_{1i} = \{l_{jk} | \exists link(t_j \rightarrow t_k), t_j \in S_{1i}\};$$

(4) Constructing the data set $S_{2i}, i = [1, 6])$

$$S_{2i} = \{t_n | link(t_m \rightarrow t_n) \in L_{1i} \wedge (t_m \in S_{1i}) \wedge (t_n \notin S_{1i})\}.$$

(5) Combining S_{1i} and S_{2i}.

$$S_i = S_{1i} \cup S_{2i}, i \in [1, 6].$$

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶

Go Back
Full Screen
Close
Quit

3.3. Eigenvalue

- Irreducible

  As $T = q \cup S_1 \cup S_2$,

  - $\exists$ Link $q_i \longrightarrow t_i, t_i \in S_1$;
  - $\exists$ Link $t_i \longrightarrow q_i, t_i \in S_1$;
  - $\exists$ Link $t_i \longrightarrow s_j, t_i \in S_1 \wedge s_j \in S_2$;
  - $\exists$ Link $s_j \longrightarrow q_i, s_j \in S_2$;

  That means we can reach each other pages from one of the pages in set S.

- Aperiodic

  In the matrix T, all elements in the diagonal are positive except the $t_{00} = 0$. $T_{ii} > 0, (i \in (0, N])$ is always true.

T has a real, positive, and the biggest eigenvalue.

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 21 of 31
Go Back
Full Screen
Close
Quit

3.4. Experiment Results

- ρ_2

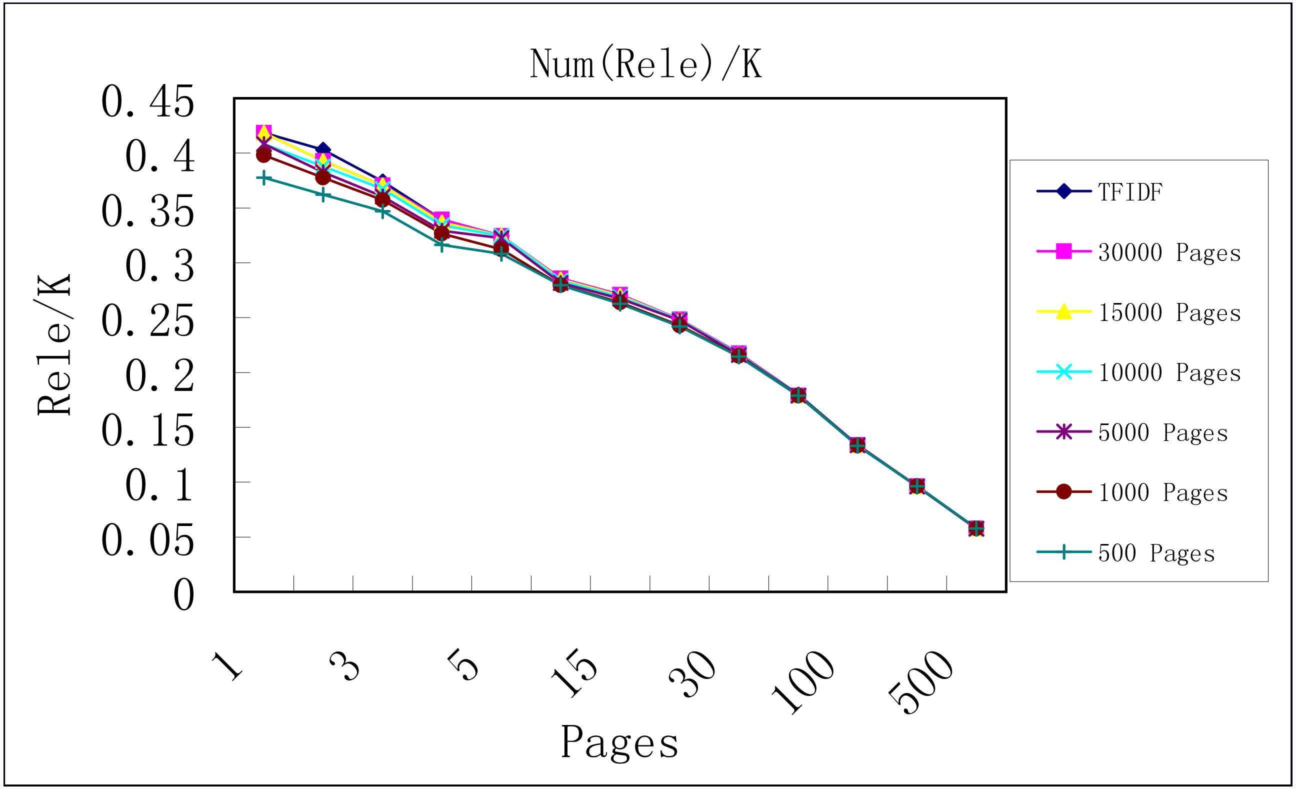
	S1	S2	Precision	Recall	\rho_2
Q1	25767	113119	0.000776	0.909091	0.000012
Q2	30000	55940	0.006933	0.773234	0.00013
Q3	4390	20952	0.022096	0.941748	0.000061
Q4	30000	50399	0.0043	0.921429	0.000081
Q5	30000	52262	0.000733	0.916667	0.000014
Q6	30000	52026	0.0003	0.642857	0.000006
Q7	21126	78488	0.002887	0.884058	0.000038
Q8	30000	68911	0.000633	0.542857	0.000012
...	...	...	...	...	...
Q100	30000	48004	0.001533	0.779661	0.000029
Avg			0.0052235	0.8266822	3.025E-05

$$\rho_2 = 3.025 * 10^{-5} \ll 1.$$

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 22 of 31
Go Back
Full Screen
Close
Quit

- Number(Relevance page)/ Number(top-N)



3.5. Discussing the result

Solution x_0,X_1,and X_2 of T' are depended on V, where

$$V = (I - \rho_1 * M'_{11})^{-1} * P(q).$$

If the $(\rho_1 * M'_{11})$ is small, the V and the $P(q)$ will be very similar. As the M'_{11} is the truth, the ρ_1 decides the value of V.
In our experiment, because the ρ_1 is small, the final solution is similar to the $P(q)$, the `TFIDF` value.

3.6. 需要确认的问题

(1) 用这种方法提取出来的链接组织的图能够保持 $Internet$ 的特性吗?

(2) 这种方法与 $TFIDF$ 方法相比是两种不同的方法吗?

(3) 新方法比旧方法要好吗?

解决方法:

(1) 通过测试链接密度的方法;

(2) 通过计算“Spearman Rank Correlation Coefficient”的值加以确定;

(3) 采用计算 $Precision$ 的办法加以确定。

4 Conclusion and Future work

4.1. Conclusion

- The methods considering the hyper-link or the content solely have shortages, such as the quantity of computation and the precision of retrieval, etc.
- The result shows that the precision of new method approaches the `TFIDF`'s. But the new framework has less quantity of computation than `TFIDF`.
- We should get another data set for test. By changing the parameter of ρ_1 and ρ_2 to observe the results.
- About the relation of Content-based and Link-based, it is not enough to take account of the simple links(URL), it may be better to consider the semantic links among the pages.

4.2. Future Works

(1) Distribution: By using $TFIDF$ to measure the probability from the query Q to the pages, the results is not ideal, is there any other way to *instead of it*?

(2) Test Page Set: By using the WT10g to measure the feedback pages according to the Q, the Num(relevance)/K is not better than $TFIDF$ does as the less total relevance pages. Is there any other set to *instead of it*?

(3) Is there any new model to describe the relationship between the query and the pages set?

(4) In another domain, how to retrieval the graphic file?

5 Appendix

5.1. Measure of Relevance

$$Rele(Q,d_i)=Distance(Q,d_i).$$

5.2. Description of Probability distribution

用户从查询 Q 转移到页面 d_i 的概率为序列：$(p_1(q),p_2(q),\cdots,p_n(q))$，概率$p_i(q)$可以由多种方法加以描述，最直接的方法就是：

$$p_i(q)=\frac{Rele(Q,d_i)}{\sum\limits_j(Rele(Q,d_j))}.$$

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 28 of 31
Go Back
Full Screen
Close
Quit

也可以通过下面的方法将它定义为条件概率：

假设 $p_{ij}(q), \forall i, j, q$ 用来表示页面 d_i 被检索系统认为是与查询 Q 相关的，同时存在链接 $link(d_i \rightarrow d_j)$，在此条件下 d_j 被检索系统认为是与查询 Q 相关的概率。显然：

$$\sum_j p_{ij}(q) = 1 \ \ \forall i, q.$$

假设对于页面 d_i 有 n 个对外的链接，标志这些链接为: $l_i, i \in [1, n]$，则:

- 考虑链接的权重

  假设每个链接 $l_i, i \in [1, n]$ 对应着各自的权重 $m_i, i \in [1, n]$，则定义:

$$p_{ij} = \frac{m_{ij}}{\sum_j m_{ij}}, \quad \forall i \wedge i \neq j.$$

- 不考虑链接的权重

  $m_i, i \in [1, n]$ 只有两个取值 0 和 1，则定义:

$$p_{ij} = \frac{\delta_{ij}}{\sum_j \delta_{ij}}, \quad \text{其中：} \ \delta_{ij} = \begin{cases} 1 & \exists\, Link(d_i \longrightarrow d_j) \\ 0 & Otherwise. \end{cases}$$

Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 29 of 31
Go Back
Full Screen
Close
Quit

- 考虑 d_i 与 d_j 的相似度

  定义 $s(d_i, d_j)$ 为页面 d_i 与 d_j 的相似度，则定义：

$$p_{ij} = \frac{\delta_{ij}}{\sum\limits_{j} \delta_{ij}}.$$

其中：$$\delta_{ij} = \begin{cases} s(d_i, d_j) & \dfrac{s(d_i, d_j)}{\underbrace{MAX}_{j}\, s(d_i, d_j)} \geq \varphi \; \varphi \in [0, 1] \\ 0 & Otherwise. \end{cases}$$

Introduction
Content and Link Based . . .
Experimental for CLBCRA
Conclusion and Future work
Appendix
Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 30 of 31
Go Back
Full Screen
Close
Quit

Thank you !

Introduction

Content and Link Based . . .

Experimental for CLBCRA

Conclusion and Future work

Appendix

Home Page

Title Page

◀◀ ▶▶

◀ ▶

Page 31 of 31

Go Back

Full Screen

Close

Quit