

A Comparison of Implicit and Explicit Links for Web Page Classification

Dou Shen¹ Jian-Tao Sun² Qiang Yang¹ Zheng Chen²

¹Department of Computer Science and Engineering
The Hong Kong University of Science and Technology, Hong Kong

²Microsoft Research Asia, China



Outline

- Introduction
- Related Work
- Implicit and Explicit Links
- Links for Classification
- Experiments
- Conclusion and Future Work



Introduction

- Why we need Web page classification?
 - Organize the growing amount of pages
 - Facilitate other text mining applications
- How to classify Web pages?
 - Classification algorithm (SVM, NB, KNN...)
 - Web page representation



Introduction (Cont.)

- Web page representation
 - Content Based
 - Utilize words or phrases of a target page
 - However, very often a Web page contains enough textual clues
 - Context Based
 - Leverage hyperlinks to connect pages
 - It works. However, the hyperlinks sometimes may not reflect true relationships in content between Web pages
 - **Any other kind of linkages can be defined and used?**
 - **How to improve classification with the new links?**



Related Work

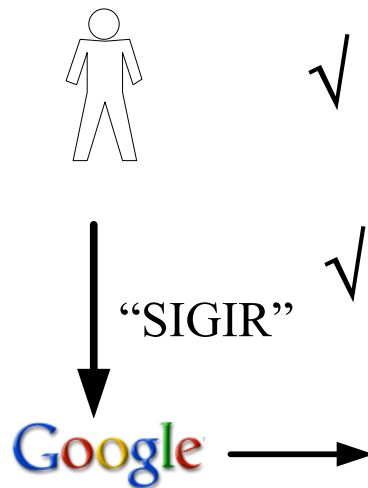
■ Exploiting Hyperlinks

- Chakrabarti et al. used predicted labels of neighboring documents to reinforce classification decisions for a given document;
- Furnkranz also reported a significant improvement in classification accuracy when using the link-based method as opposed to the full-text alone.

■ Exploiting Query Logs

- Beeferman and Berger proposed an innovative query clustering method based on query log;
- Xue et al. proposed a novel categorization algorithm named IRC to categorize the interrelated Web objects by leveraging query log.

Implicit and Explicit Links



■ Query logs

$Entry := \langle U, Q, D_q, T \rangle$

[ACM SIGIR Special Interest Group on Information Retrieval Home](#)
ACM SIGIR addresses issues ranging from theory to user demands in the the acquisition, ... The Digital Symposium Collection DVD-ROM is now a an additional \$10 per year. ...

www.sigir.org/ - 7k - [网页快照](#) - [类似网页](#)

[SIGIR 2006—Seattle](#) - [[翻译此页](#) [BETA](#)]

Space Needle SIGIR is the major international forum for the presentation of the ... The 29th Annual International ACM SIGIR Conference will be held in Washington Campus in Seattle, WA, ...

www.sigir2006.org/ - 8k - [网页快照](#) - [类似网页](#)

[Departamento de Ciência da Computação](#)

Bem-Vindo ao Webmail do DCC. Prompt de Login. Login: Password: Qu - 10:57.

webmail2.dcc.ufmg.br/ - 4k - [网页快照](#) - [类似网页](#)

[Special Inspector General for Iraq Reconstruction : SIGIR Home](#)

SIGIR, the successor to the Coalition Provisional Authority Inspector General by Congress to ... SIGIR releases its first report of the Lessons Learned In ongoing efforts in Iraq, ...

www.sigir.mil/ - 14k - [网页快照](#) - [类似网页](#)

[SIGIR 2003](#) - [[翻译此页](#) [BETA](#)]

SIGIR is the major international forum for the presentation of new research.



Implicit and Explicit Links (Cont.)

- Implicit link 1 (L_{11})
 - **Assumption:** a user tends to click the pages related to the issued query;
 - **Definition:** there is an L_{11} between d_1 and d_2 if they are clicked by the same person through the same query;
- Implicit link 2 (L_{12})
 - **Assumption:** users tend to click related pages according to the same query
 - **Definition:** there is an L_{12} between d_1 and d_2 if they are clicked according to the same query



Implicit and Explicit Links (Cont.)

- Comparison between L_1 and L_2
 - The constraint of L_2 is not as strict as that for L_1 ;
 - Thus, there are more links of L_2 can be constructed than L_1 ;
 - L_2 is noisier than L_1 , especially for the ambiguous queries (such as “apple”)



Implicit and Explicit Links (Cont.)

- Three kinds of Explicit Links defined based on hyperlinks

$$L_E(i, j) = \begin{cases} 1 & \text{Cond}_E \\ 0 & \text{Other} \end{cases}$$

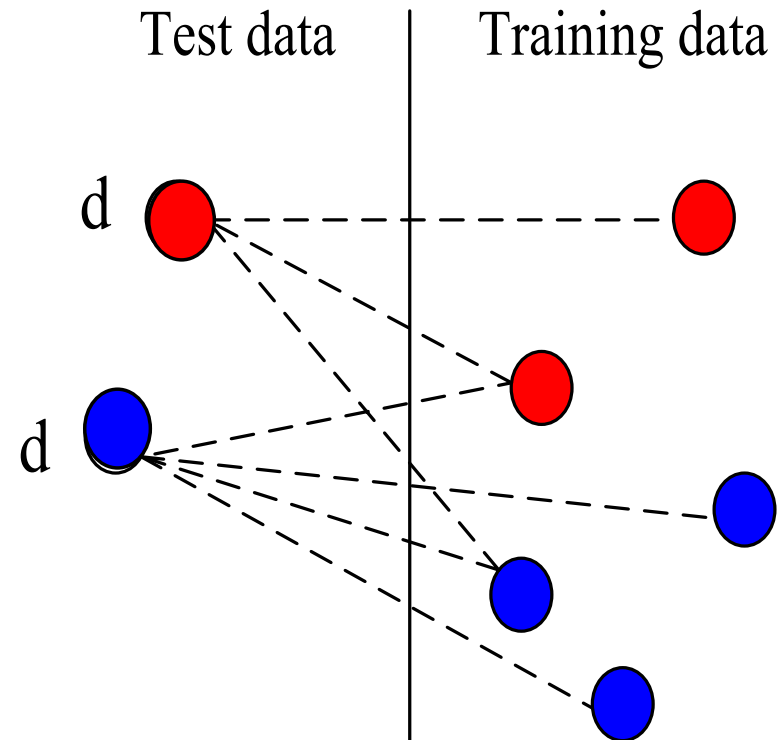
- *Cond_E1*: there exists hyperlinks from d_j to d_i , (In-Link to d_i from d_j)
- *Cond_E2*: there exists hyperlinks from d_i to d_j , (Out-Link from d_i to d_j)
- *Cond_E3*: either *Cond_E1* or *Cond_E2* holds

Links for Classification

- Classification by Linking Neighbors (CLN)

$$\text{Category}(d) = \underset{c_i}{\operatorname{argmax}} \left(\text{count}_{d_j \in S_L(d)} (d_j \in c_i) \right)$$

- CLN is similar to KNN;
- K is not a constant as in KNN and it is decided by the set of the neighbors of the target page.





Links for Classification (Cont.)

- Build Virtual Document

Given a document, the virtual document is constructed by borrowing some **Extra Text** from its neighbors

- Extra Text

- Local Text: Plain text + Meta Data
- Anchor Text
- Extended Anchor Text
- Anchor Sentence

- Apply any classifier such as SVM, NB



Links for Classification (Cont.)

- Local Text:
 - Plain text: remaining text by removing html tags;
 - Meta Data: text between <Meta> and </Meta>;
- Anchor Text
 - The visible text in a hyperlink
- Extended Anchor Text
 - The set of rendered words occurring up to 25 words before and after an associated link
- Anchor Sentence
 - The set of sentences containing the query based on which the implicit link is created



Experiments

- Datasets
 - 1.3 million Web pages among 424 classes from Open Directory Project (ODP)
 - 44.7 million records in 29 days from MSN
- Classifiers
 - Naïve Bayesian Classifier;
 - Support Vector Machine (SVM^{light})
- Evaluation Metrics
 - Precision, Recall, F1



Experiments (Cont.)

- Statistics of Links

Link type [↵]	Consistency [↵]	#Links [↵]	#Link/page [↵]
L _I 1 [↵]	0.569 [↵]	162901 [↵]	2.00 [↵]
L _I 2 [↵]	0.462 [↵]	484462 [↵]	4.63 [↵]
L _E 1 [↵]	0.458 [↵]	1148217 [↵]	4.38 [↵]
L _E 2 [↵]	0.458 [↵]	1148217 [↵]	3.23 [↵]
L _E 3 [↵]	0.437 [↵]	1056208 [↵]	2.18 [↵]

- $\#L_E1 = \#L_E2 > \#L_E3$

- $A \rightarrow B; B \rightarrow C; C \rightarrow B$

- $\#L_E1 = 3; \#L_E2 = 3; \#L_E3 = 2$

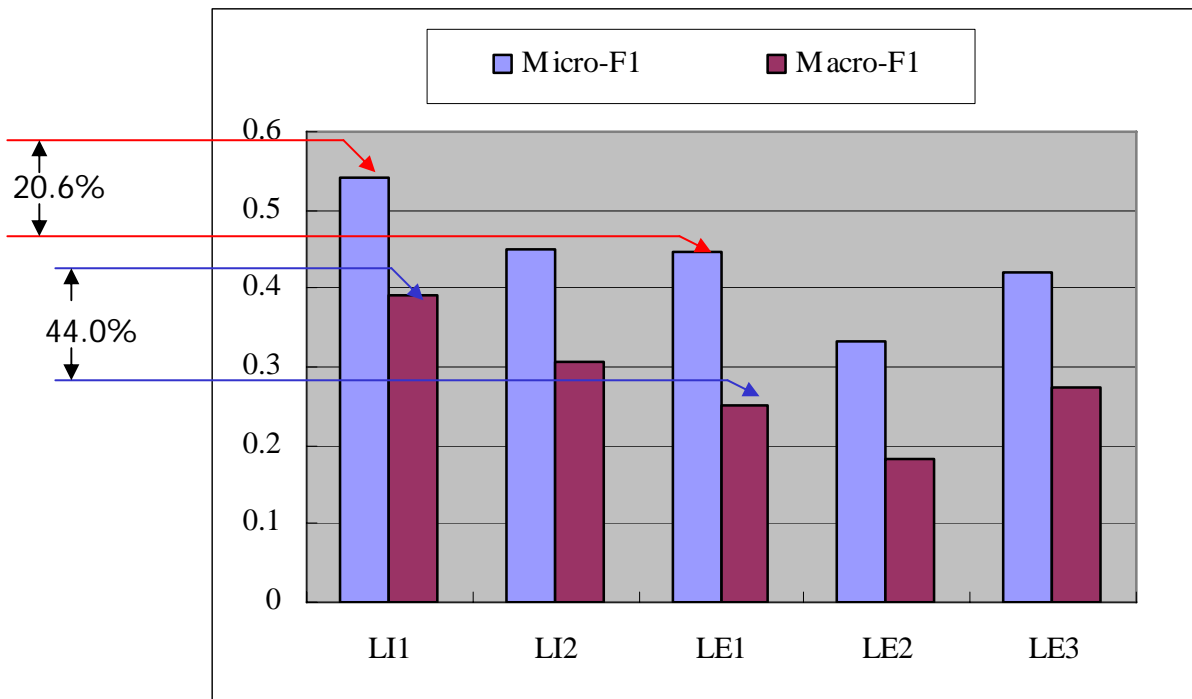
- Consistency:

the percentage of links that have the two linked pages from the same category.

- The consistency of L_I1 is much higher than others;
- The consistency values of all explicit links are lower than 50%, which explained some published results that it is not helpful to use hyperlink in a straightforward way;

Experiments (Cont.)

■ Results of CLN on Different Links

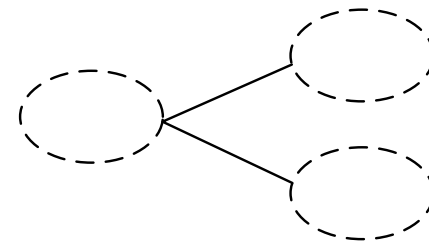
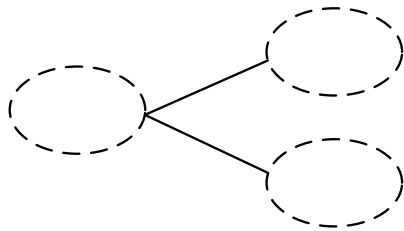


- The results are consistent with the consistency values of different kinds of links
- Compare the best result of implicit links and the best result of explicit links

Experiments (Cont.)

- Construction of virtual documents

φ	Explicit Link(L_E1) φ	Implicit Link(L_I1) φ
$LT\varphi$	$ELT\varphi$	$ILT\varphi$
$AS, EAT, AT\varphi$	$AT, EAT\varphi$	$AS\varphi$





Experiments (Cont.)

- Performance on different kinds of VD

(1) Classification Performance by SVM

	LT	ILT	ELT	AS	EAT	AT
Mi-F1	0.607	0.652	0.629	0.591	0.519	0.403
Ma-F1	0.348	0.444	0.384	0.389	0.297	0.253

(2) Classification Performance by NB

	LT	ILT	ELT	AS	EAT	AT
Mi-F1	0.551	0.583	0.515	0.556	0.464	0.361
Ma-F1	0.25	0.336	0.298	0.296	0.226	0.163

- The performance of AS, EAT and AT is just as good as the baseline, or even worse.
- ILT is much better than ELT
- ELT is better than LT, but not always



Experiments (Cont.)

- Explanation

- the average size of the virtual documents (in terms of KB)

	LT	ILT	ÉLT	AS	EAT	AT
Ave size	5.60	11.10	23.80	2.30	0.30	0.20

- the consistency or purity of the content of the virtual documents

Experiments (Cont.)

■ Effect of Different Combinations

(1) Results on SVM

	1:0	4:1	3:1	2:1	1:1	1:2	1:3	1:4	0:1	Impr*
Micro-F1										
AS	0.607	0.654	0.659	0.661	0.662	0.661	0.657	0.650	0.591	9.06%
EAT		0.616	0.627	0.636	0.638	0.639	0.635	0.628	0.519	5.27%
AT		0.596	0.606	0.612	0.611	0.612	0.615	0.610	0.403	1.30%
Macro-F1										
AS	0.348	0.423	0.420	0.432	0.426	0.431	0.429	0.422	0.389	24.14%
EAT		0.370	0.383	0.391	0.384	0.386	0.387	0.376	0.297	12.37%
AT		0.355	0.365	0.365	0.356	0.356	0.352	0.352	0.253	4.83%

(2) Results on NB

	1:0	4:1	3:1	2:1	1:1	1:2	1:3	1:4	0:1	Impr*
Micro-F1										
AS	0.551	0.626	0.627	0.629	0.622	0.618	0.613	0.612	0.556	14.16%
EAT		0.614	0.620	0.614	0.600	0.614	0.616	0.613	0.464	12.52%
AT		0.594	0.596	0.594	0.579	0.585	0.586	0.584	0.361	8.17%
Macro-F1										
AS	0.250	0.353	0.352	0.349	0.337	0.338	0.338	0.346	0.296	41.20%
EAT		0.313	0.306	0.306	0.281	0.297	0.304	0.310	0.226	25.20%
AT		0.304	0.296	0.291	0.275	0.281	0.287	0.285	0.163	21.60%



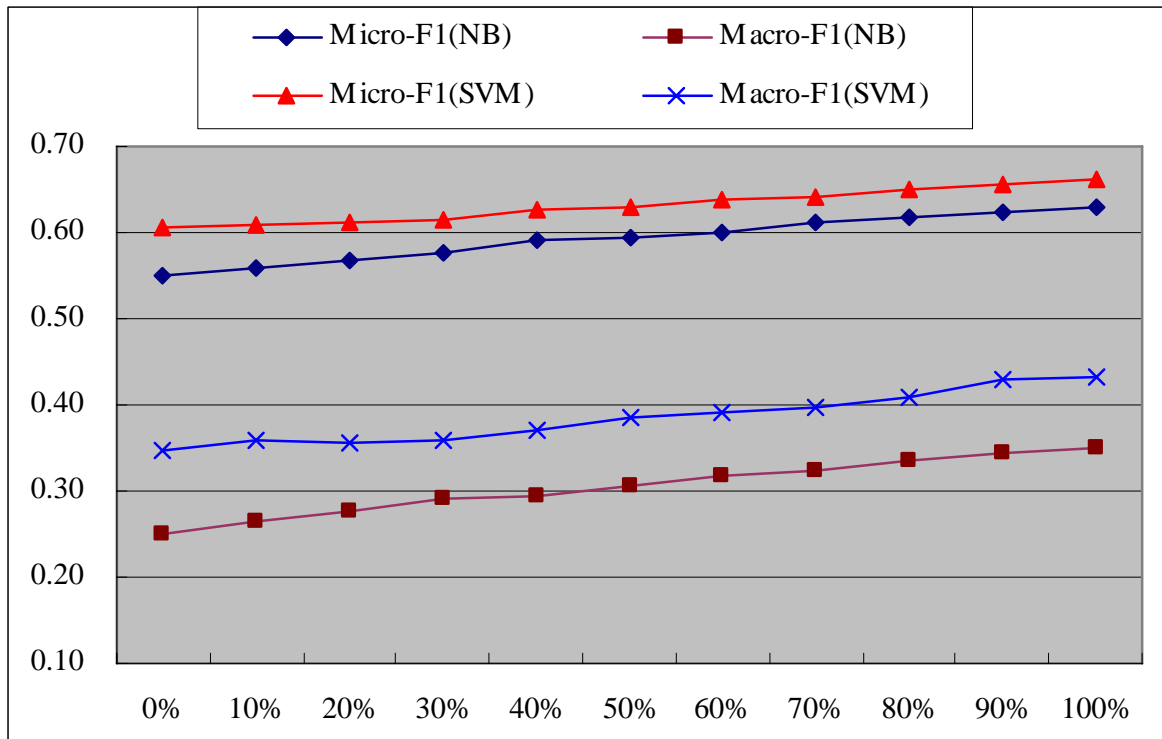
Experiments (Cont.)

- Observations

- Either AT, EAT or AS can improve the performance of classification;
- AS achieves greatest improvement;
- Different weighting schemes do not make too much of a difference
- We also tried to combine LT, EAT and AS together, no further improvement is obtained

Experiments (Cont.)

- The effect of Query Log quantity





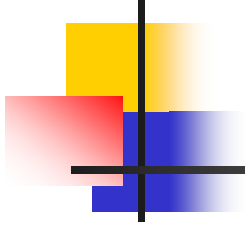
Conclusion

- Based on the query logs, a new kind of links-- the implicit links -- is introduced;
- Comparison between the implicit and explicit links on a large dataset is given;
- A concept of a virtual document by extracting "anchor sentence (AS)" through implicit links is presented;
- Experiment result show that implicit link is better than explicit when used for web page classification.



Future Work

- Introduce more kinds of implicit and explicit links;
- Try on more applications such as clustering and summarization;
- Extract other information such as “Dissimilarity Relationship” from query log.



Thanks