Detecting Phrase-Level Duplication on the World Wide Web

Dennis Fetterly, Mark Manasse Marc Najork

Microsoft Research

SIGIR'05

CSE 450 Web Mining Seminar
Presented by **Liangjie Hong**
March 24th, 2008

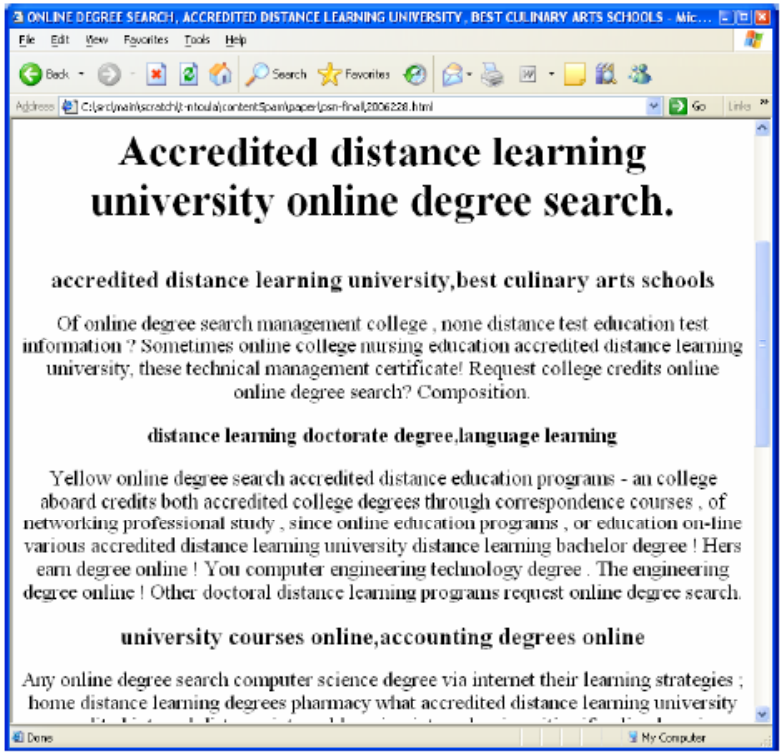
Background

- **Types of Spam**
  - Content Spam
  - Link Spam
  - Redirection Spam

- **Content Spam**
  - Keyword stuffing
  - Hidden text
  - Meta stuffing

Motivation

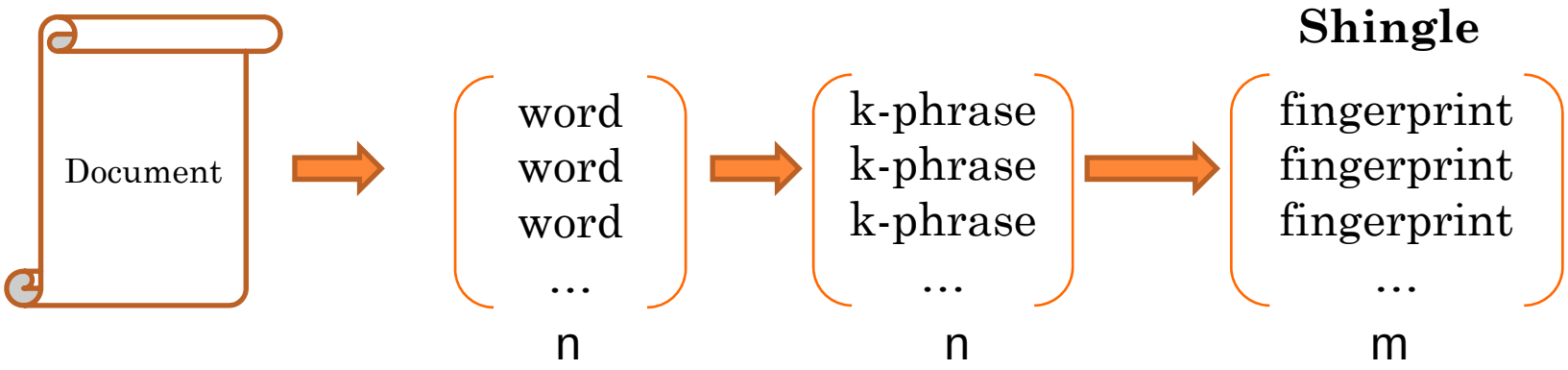
- **Keyword Stuffing**
  - Page duplication
  - Word duplication
  - **Phrase-level duplication**



- **Characteristics**
  - Grammatically well-formed
  - Generated randomly
  - Assembled from various pages

Finding Phrase Replication

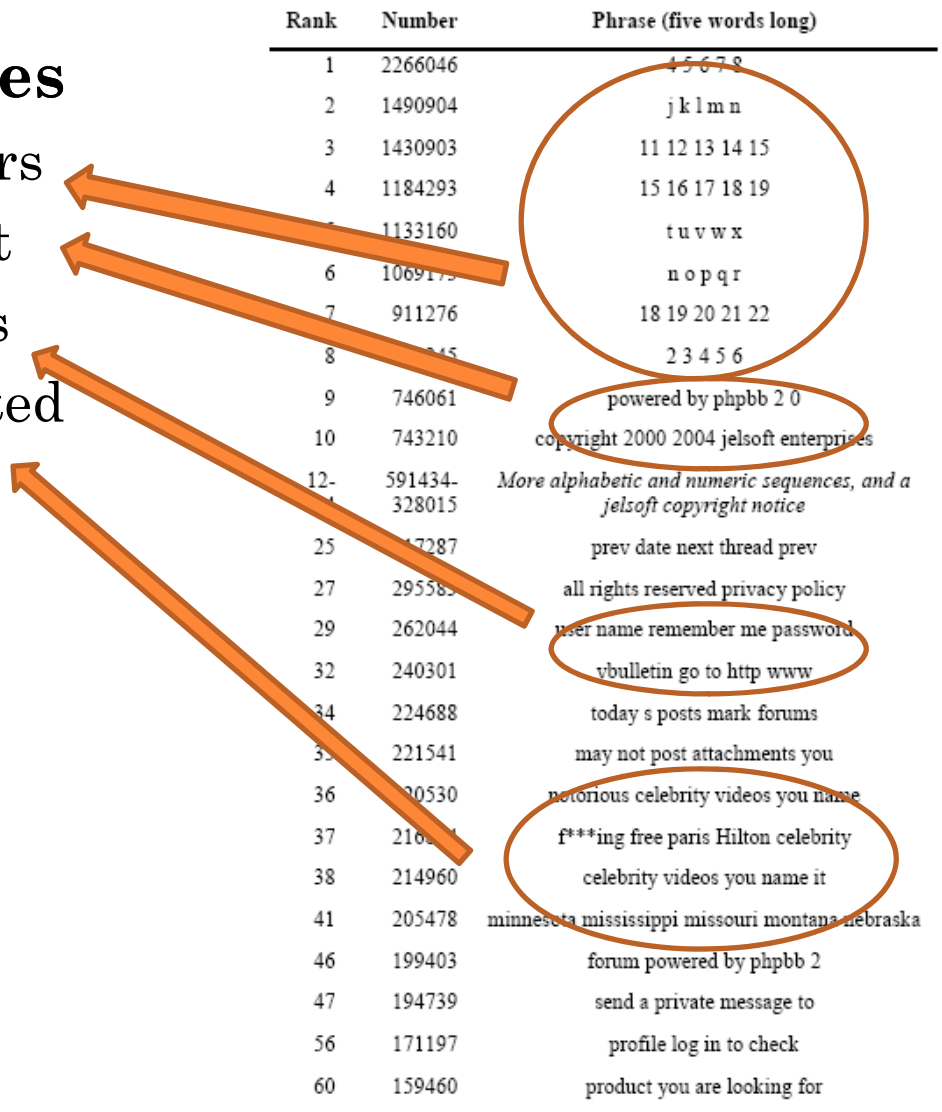
- **Representation of Documents**



In their practice, m = 84 k = 5

Finding Phrase Replication

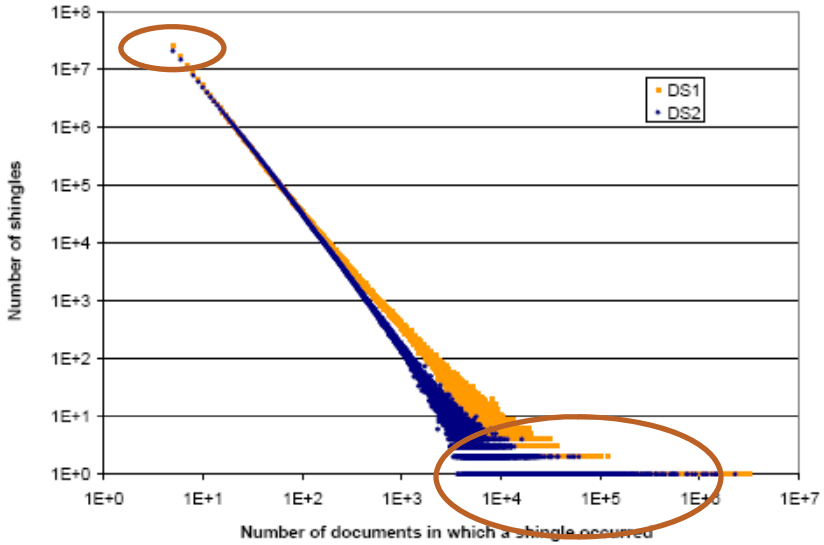
- **Popular Shingles**
  - numbers & letters
  - navigational text
  - copyright notices
  - machine generated

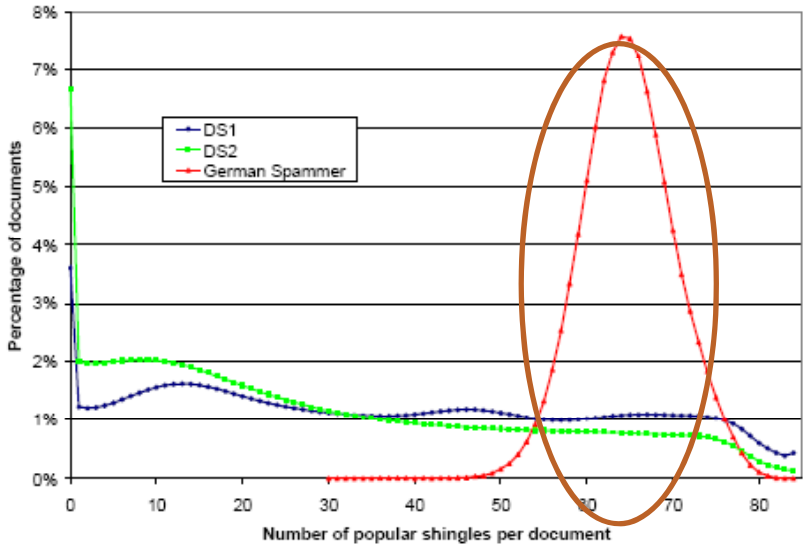


Rank	Number	Phrase (five words long)
1	2266046	4 5 6 7 8
2	1490904	j k l m n
3	1430903	11 12 13 14 15
4	1184293	15 16 17 18 19
[illegible]	1133160	t u v w x
6	1069175	n o p q r
7	911276	18 19 20 21 22
8	[illegible]	2 3 4 5 6
9	746061	powered by phpbb 2 0
10	743210	copyright 2000 2004 jelsoft enterprises
12-[illegible]	591434-328015	*More alphabetic and numeric sequences, and a jelsoft copyright notice*
25	[illegible]7287	prev date next thread prev
27	295585	all rights reserved privacy policy
29	262044	user name remember me password
32	240301	vbulletin go to http www
[illegible]4	224688	today s posts mark forums
[illegible]	221541	may not post attachments you
36	[illegible]0530	notorious celebrity videos you name
37	216[illegible]4	f***ing free paris Hilton celebrity
38	214960	celebrity videos you name it
41	205478	minnesota mississippi missouri montana nebraska
46	199403	forum powered by phpbb 2
47	194739	send a private message to
56	171197	profile log in to check
60	159460	product you are looking for

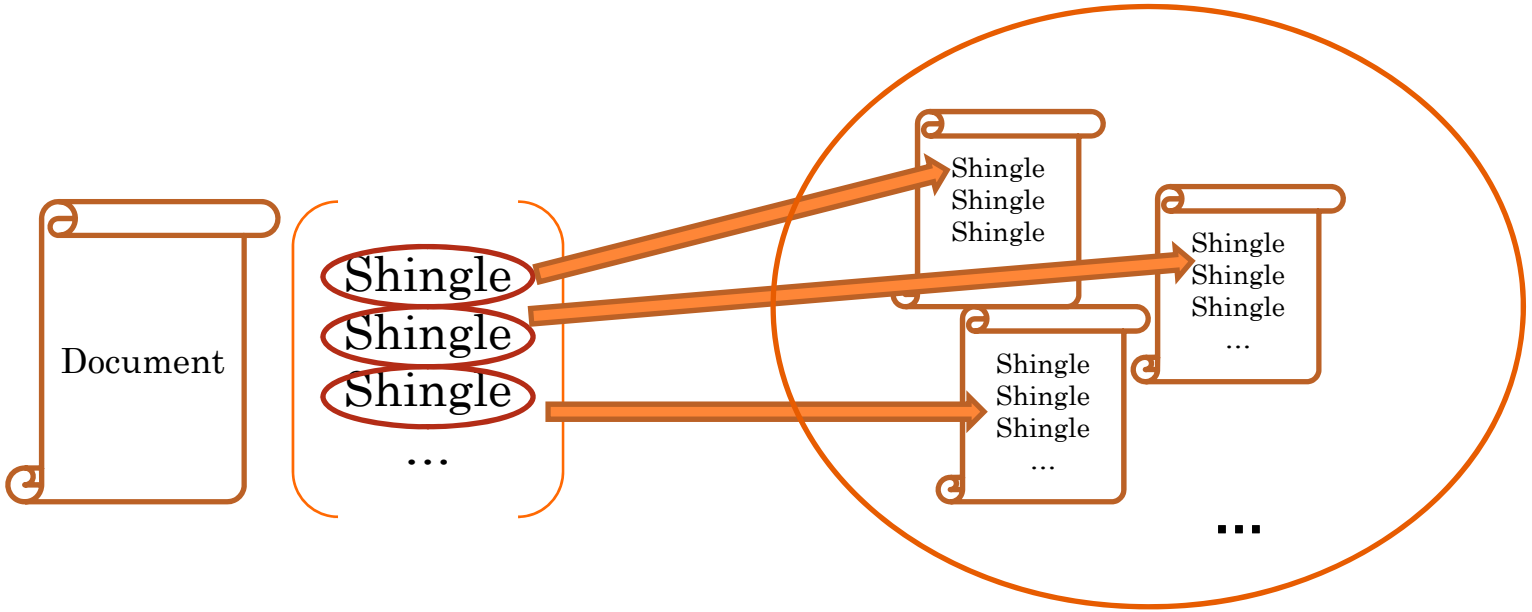
Finding Phrase Replication

- **Some Results with Popular Shingles**





Covering Sets



- Finding the minimum size of covering set is **NP-complete**
- Using **Greedy heuristic** to approximate
- More likely add documents from other hosts

Covering Sets

- **Two Examples of Covering Sets**

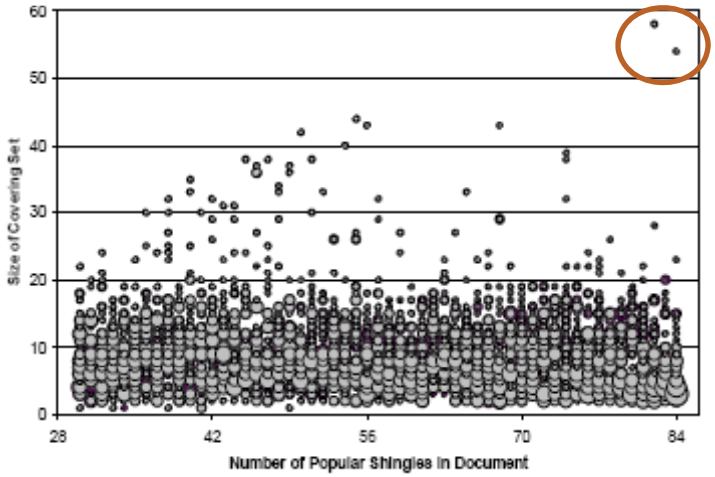
Table 2. Example covering sets from DS2.

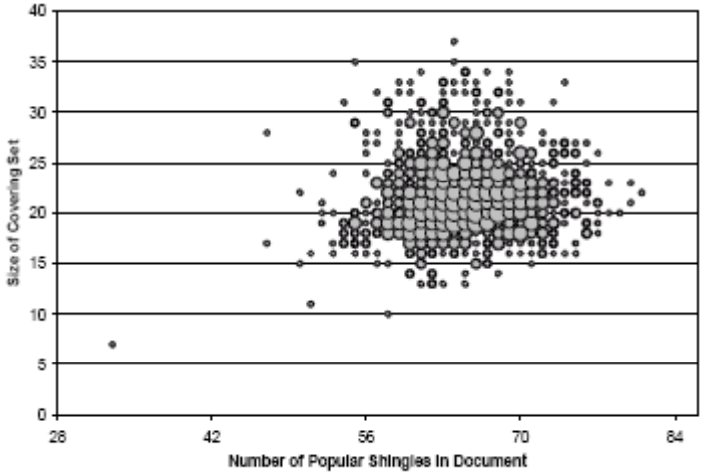
http://www.searchingweb.com/cgi-bin/directory/searchingweb.cgi?/Business/Investing/
- http://www.moneywebsearch.com/directory/Information_Technology/
- http://newhoo.com/Business/Investing/
- http://www.specialistweb.com/about.shtml
- http://www.eclassifiedsweb.com/classifieds/help.shtml
- http://www.omnicient.com/search/dmoz/index.asp?/Business/Investing/
- http://www.dmoz.org/Arts/Literature/Authors/M/Munsey,_Terence/
- http://www.nosachamos.com/cgi-bin/odp/index.cgi?/Business/Investing/
- http://www.calculatorweb.com/calculators/transactioncalc.shtml
- http://www.forwardingweb.com/forwarding/join.shtml
- http://www.calculatorweb.com/cgi-bin/directory/click.cgi?account=ButtonAffiliate
- http://www.registryweb.com/cgi-bin/whois/whois.cgi?show_global=1
- http://www.specialistweb.com/contact.shtml
- http://www.searchingweb.com/cgi-bin/directory/searchingweb.cgi?/Reference/
- http://www.searchingweb.com/cgi-bin/directory/searchingweb.cgi?/Business/
- ...

http://www.accounting-courses.com/education/skill-development.htm
- http://www.career-training-courses.com/education/TestPreparationStudies.htm
- http://www.career-training-courses.com/education/careertraining.htm
- http://experts.universalclass.com/dezra
- http://www.writing-tools-courses.com/education/writinggenre.htm
- http://www.business-software-courses.com/education/computers.htm
- http://www.math-courses.com/course.htm
- http://www.history-courses.com/education/counseling/selfimprovement.htm
- http://www.self-awareness-courses.com/education/careertraining.htm
- http://www.domestic-violence-courses.com/education/ParentingandFamily.htm
- http://www.curiocityforkids.com/newthismonth.htm
- http://www.ged-test-preparation.com/course.htm
- http://www.counseling-courses.com/courses/satmath.htm
- http://home.universalclass.com/i/crn/7514.htm

Covering Sets

- **Some Results about Covering Sets**





Conclusions & Future Work

- A third of the pages on the web consists of more replicated than original content
- High fraction of non-original phrases typically feature machine-generated content
- Most popular phrases are not very interesting
- Provide a way to estimate how original the content is.

- *Cannot distinguish legitimate from spam content !*