

# Site Level Noise Removal for Search Engines



**André Luiz da Costa Carvalho**

*Federal University of Amazonas, Brazil*

**Paul-Alexandru Chirita**

*L3S and University of Hannover, Germany*

**Edleno Silva de Moura**

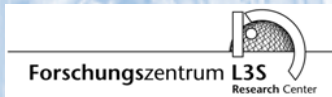
*Federal University of Amazonas, Brazil*

**Pável Calado**

*IST/INESC-ID, Portugal*

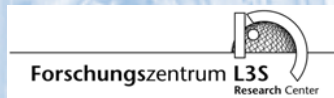
**Wolfgang Nejdl**

*L3S and University of Hannover, Germany*



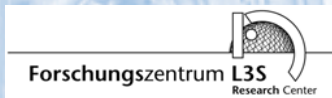
# Outline

- Introduction
- Proposed Noise Removal Techniques
- Experiments
- Practical Issues
- Conclusion and future work



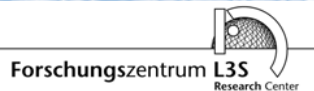
# Introduction

- Link analysis algorithms are a popular source of evidence for search engines.
- These algorithms analyze the Web's link structure to assess the quality (or popularity) of web pages.



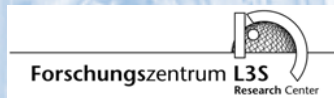
# Introduction

- This strategy relies on considering links as votes for quality.
- But not every link is a true vote for quality.
- We call these links “noisy links”



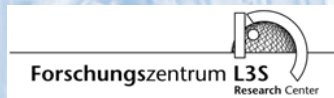
# Examples

- Link Exchanges between friends;
- Tightly Knit Communities;
- Navigational links;
- Links between mirrored sites;
- Web Rings;
- SPAM.



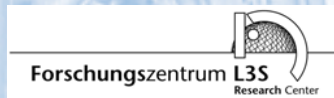
# Introduction

- In this work we propose methods to identify noisy links.
- We also evaluate the impact of the removal of the identified links.



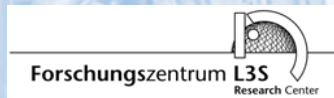
# Introduction

- Most of the previous works are focused on SPAM.
- We have a broader focus, focusing on all links that can be considered noisy.
- This broader focus allow our methods to have a greater impact on the database.



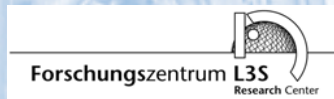
# Introduction

- In this work, we propose site level analysis based methods, i.e., methods based on the relationships between sites instead of pages.
- Site Level Analysis can lead to new sources of evidence, that aren't present on page level.
- Previous works are solely based on page level analysis.



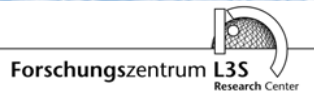
# Proposed Noise Removal Techniques

- Uni-Directional Mutual Site Reinforcement (UMSR);
- Bi-Directional Mutual Site Reinforcement (BMSR);
- Site Level Abnormal Support (SLAbS);
- Site Level Link Alliances (SLLA);



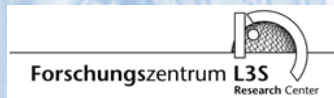


# Site Level Mutual Reinforcement



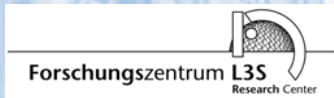
# Site Level Mutual Reinforcement

- Based on how connected is a pair of sites.
- Assumption:
  - *Sites that have many links between themselves have a suspicious relationship.*
- Ex: Mirror Sites, Colleagues, Sites from the same group.

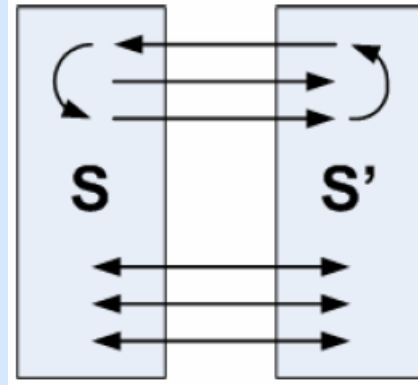


# Uni-Directional and Bi-Directional

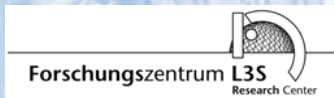
- Uni-Directional
  - Counts the number of links between the sites.
- Bi-Directional
  - Counts the number of link exchanges between pages of the sites.



# Site Level Mutual Reinforcement

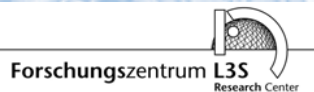


- In this example, we have 3 link exchanges, and a total of 9 links within this pair of sites.



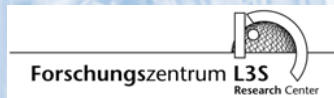
# Site Level Mutual Reinforcement

- After counting, We remove all links between pairs that have more links counted than a given threshold.
- This threshold was set by experiments.



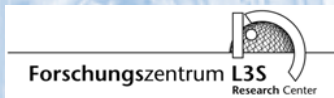


# Site Level Abnormal Support



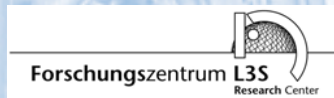
# Site Level Abnormal Support

- Based on the following assumption:
  - *The total amount of links to a site (i.e., the sum of links to its pages) should not be strongly influenced by the links it receives from some other site.*
- Quality sites should be linked by many different sites.



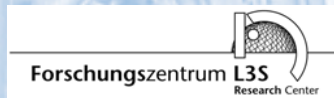
# Site Level Abnormal Support

- Instead of plain counting, we calculate the percentage of the total incoming links.
- If this percentage is higher than a threshold, we remove all links between this pair of sites.



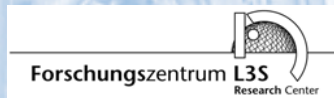
# Site Level Abnormal Support

- For example, if a site A has 100 incoming links, where 10 of that links are from B, B is responsible for 10% of the incoming links to site A.



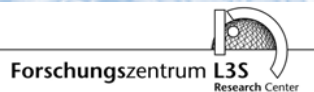
# Site Level Abnormal Support

- Using percentage avoid some problems of the plain counting of Mutual Reinforcement methods.
- For instance, tightly knit communities with sites having few links between themselves can be detected.



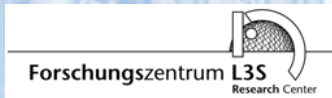


# Site Level Link Alliances



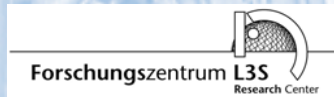
# Site Level Link Alliances

- Assumption:
  - *A Web Site is as Popular as diverse and independent are the sites that link it.*
- Sites Linked by a tight community aren't as popular as sites linked by a diverse set of sites.



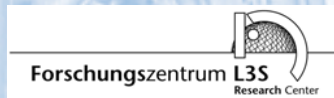
# Site Level Link Alliances

- The impact of these alliances on PageRank was previously presented on the literature, but they did not present any solution to it.

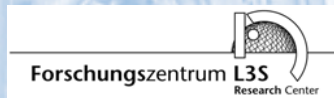
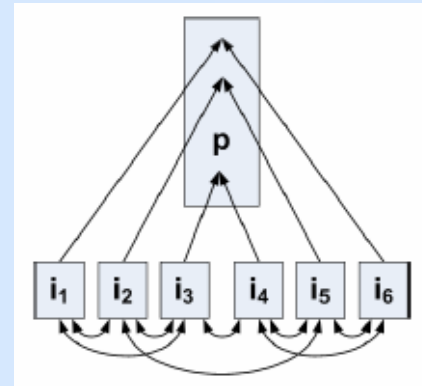


# Site Level Link Alliances

- We are interested to know, for each page, how connected are the pages that point to it, considering links between pages in different sites.
- We called this tightness “susceptivity”



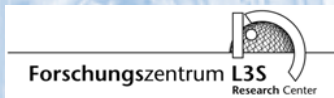
# Site Level Link Alliances



- The *Susceptivity* of a page is, given the set of pages that link to it, the percentage of the links from this set that link to others pages on the same set.

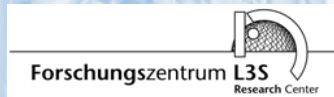
# Site Level Link Alliances

- After the calculus of the susceptibility, the incoming links of a page are downgraded with  $(1 - \text{susceptivity})$ .
- In PageRank, which was the baseline of the evaluation of the methods, this downgrade was integrated in the algorithm.



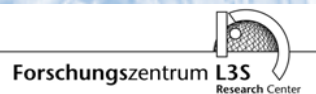
# Site Level Link Alliances

- At each iteration, the value downgraded from each link is uniformly distributed between all pages, to ensure convergence.



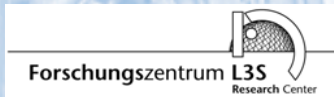


# Experiments



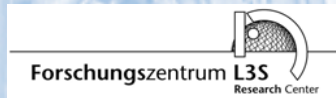
# Experiments

- Experimental Setup
  - The performance of the methods was evaluated by the gain obtained in the PageRank algorithm.
  - We used in the evaluation the database of the TodoBR search engine, a collection of 12,020,513 pages connected by 139,402,345 links.



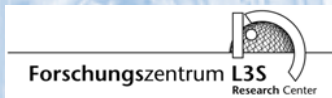
# Experiments

- Experimental Setup
  - The queries used in the evaluation were extracted from the TodoBR log, composed of 11,246,351 queries.



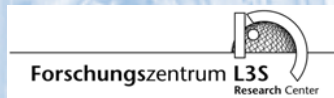
# Experiments

- **Experimental Setup**
  - We divided the selected queries in two sets:
    - **Bookmark Queries**, in which a specific Web page is sought.
    - **Topic Queries**, in which people are looking for information on a given topic, instead of some page.



# Experiments

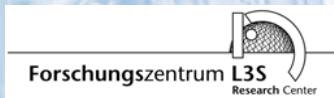
- **Experimental Setup:**
  - Each set was further divided in two subsets:
    - **Popular Queries:** The top most popular bookmark/topic queries.
    - **Randomly Selected Queries.**
  - Each subset of bookmark queries contained 50 queries, and each subset of topic queries contained 30 queries.



# Experiments

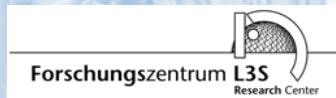
- Methodology

- For processing the queries, we selected the results where there was a Boolean match of the query, and sorted these results by their PageRank scores.
- Combinations with other evidences was tested, and led to similar results, but with the gains smoothed.



# Experiments

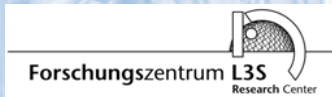
- Methodology:
  - Bookmark queries evaluation was done automatically, while topic queries evaluation was done by 14 people.
  - These people evaluated each result as *relevant* and *highly relevant*.
  - This lead to two evaluations for each query: *considering both relevant and highly relevant* and *considering only highly relevant*.



# Experiments

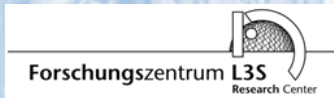
- Methodology:

- Bookmark queries were evaluated using the Mean Reciprocal Rank (MRR).
- In bookmark queries we also used the Mean Position of the right answers as a metric.



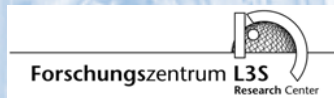
# Experiments

- Methodology:
  - For topic queries, we evaluated the Precision at 5 ( $P@5$ ), Precision at 10 ( $P@10$ ) and MAP (Mean Average Precision)



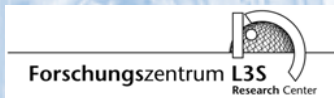
# Experiments

- Methodology:
  - We evaluated each method individually, and also evaluated all possible combinations of methods.



# Experiments

- Algorithm specific aspects:
  - The concept of site adopted in the experiments was the host part of the *URL*.
  - We adopted the MRR as a measure to determine which threshold is the best for each algorithm, being the best the following:

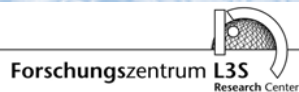


Method	Threshold
UMSR	250
BMSR	2
SLAbS	2%

# Experiments - Results

- For popular bookmark queries:

Method	MRR	Gain%	MPOS	Gain
All Links	0.3781	-	6.35	-
UMSR	0.3768	-0.55%	6.25	1.57%
SLLA	0.4241	12.14%	5	27.06%
SLLA+BMSR+SLAbS	0.4802	26.98%	4.62	37.29%



# Experiments - Results

- For random Bookmark queries:

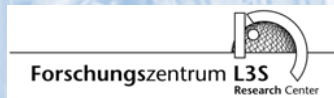
Method	MRR	Gain	MPOS	Gain
All Links	0.3200	-	8.38	-
UMSR	0.3018	-5.68%	8.61	-2.71%
SLLA	0.3610	12.88%	7.42	12.89%
SLLA+BMSR+SLAbS	0.3870	20.92%	7	19.76%



# Experiments - Results

- For popular topic queries:

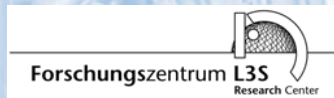
Method	MAP Highly	MAP All
All Links	0.198	0.311
UMSR	0.207*	0.333
SLLA	0.227	0.327
SLLA+BMSR+SLAbS	0.223	0.346



# Experiments - Results

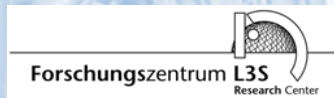
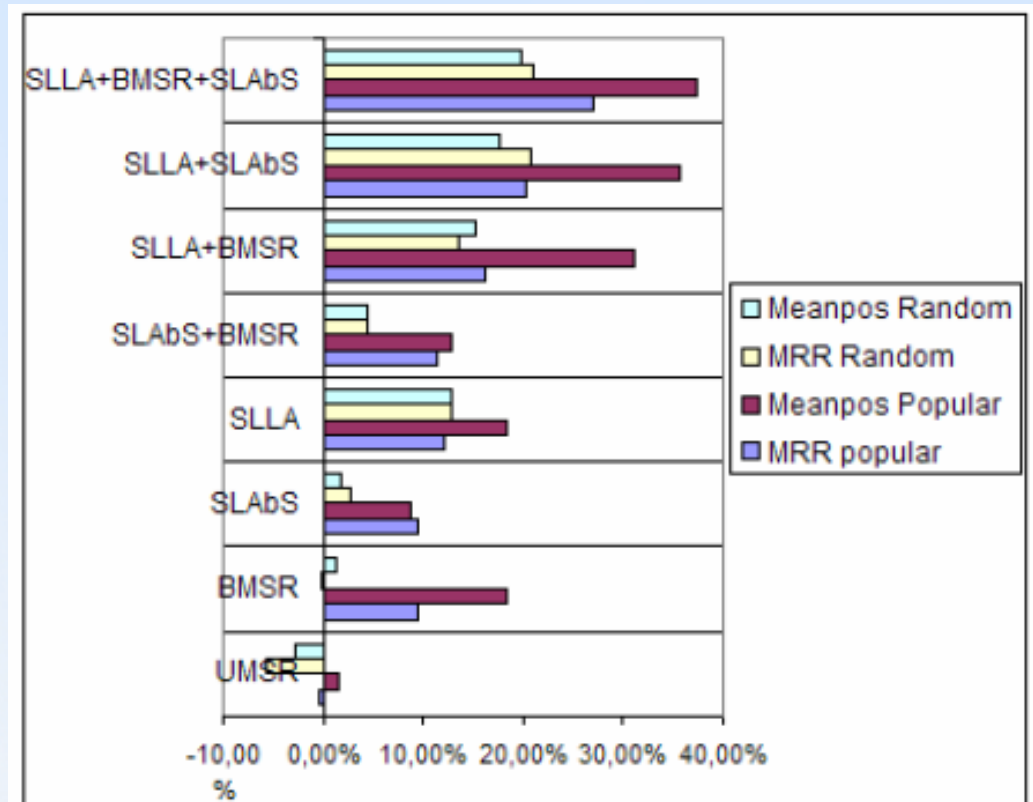
- For random topic queries:

Method	MAP Highly	MAP All
All Links	0.112	0.187
UMSR	0.131	0.196
SLLA	0.163	0.194
SLLA+BMSR+SLAbS	0.179	0.208



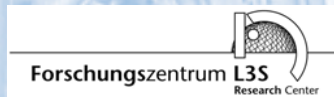
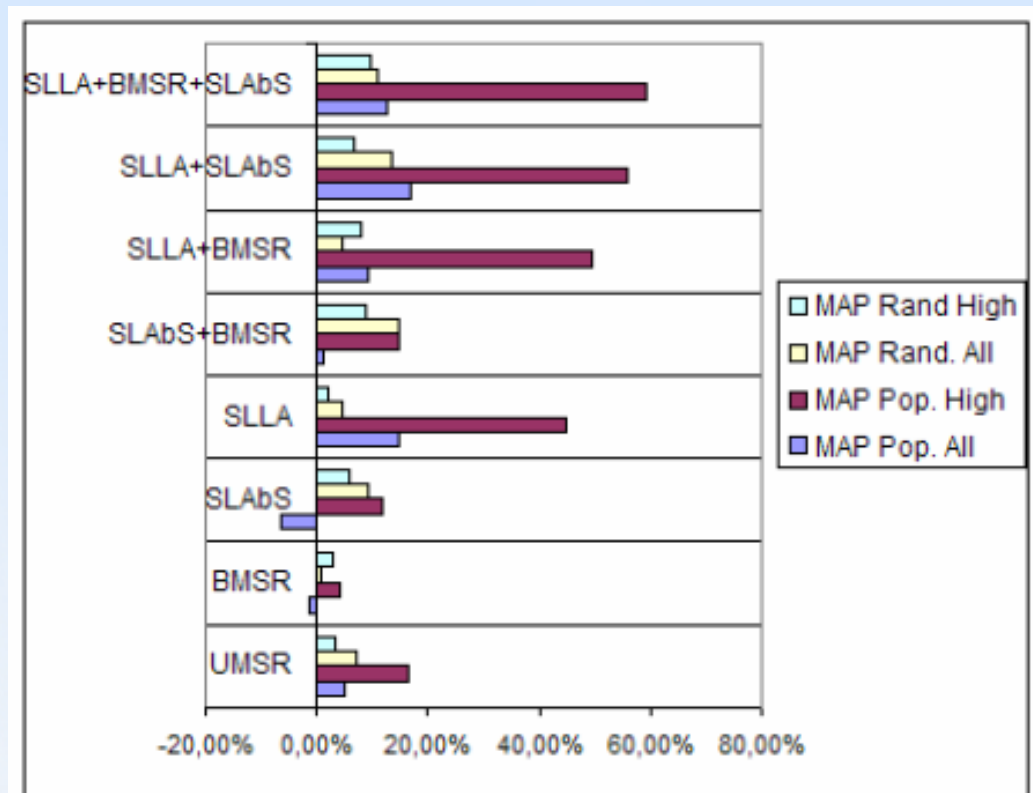
# Experiments - Results

- Relative gain for bookmark queries:



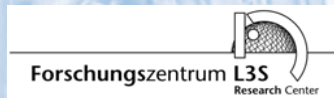
# Experiments - Results

- Relative gain for topic queries:



# Experiments

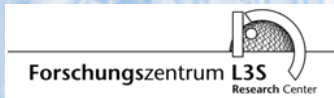
- Amount of removed links :



Method	Links Detected	% of total Links
UMSR	9371422	7.16%
BMSR	1262707	0.96%
SLAbS	21205419	16.20%
UMSR+BMSR	9507985	7.26%
BMSR+SLAbS	21802313	16.66%

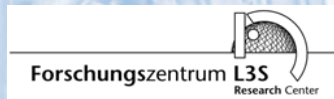
# Practical Issues

- Complexity :
  - All Proposed methods have computational cost growth proportional to the number of pages in the collection and the mean number of links per page.



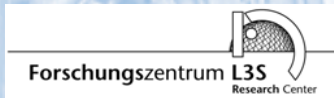
## Conclusions and Future Work

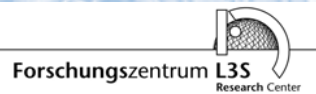
- The proposed methods obtained improvements up to 26.98% in MRR and up to 59.16% in MAP.
- Also, our algorithms identified 16.7% of the links of the database to be noisy.



# Conclusions and future work

- In future work, we'll investigate:
  - The use of different weights for the identified links instead of removing them.
  - The impact on different link analysis algorithms.





Questions ?