CS 221 Project 3 M3 - Search Engine

Jun Ma 29846938
Chang Liu 47927446
Lu Chen 35888463

1.Improvements made for Search Engine

In the milestone 2, we realized the basic functions of our search engine. In this final stage of the Search Engine project, we studied different strategies to optimize the ranking of the search results as shown below:

- Generate new Inverted Index of anchor text in both one-gram and two-gram.
- Generate new 2-gram Inverted Index of both title and body of the document.
- Optimize the ranking quality based on weigh adjustment on inverted index of document & anchor text and Page Rank.
- Automation on NDCG computation and Google search result extraction.

Based on the previous two steps and improvements above, we get the NDCG sores before and after performance improvements as shown in table 1. Obviously, our search engine results are improved in almost all queries.

Table 1. NDCG Score before and after performance improvements

Queries	NDCG before improvement	NDCG after improvement
mondego	0.61	0.61
machine learning	0.26	0.33
software engineering	0.16	0.17
security	0.23	0.19
student affairs	0.24	0.41
graduate courses	0	0.48
Crista Lopes	0	0.41
REST	0.44	0.44
computer games	0	0.28
information retrieval	0.24	0.58

2: Ranking Algorithm

- Read in the query, apply two-gram algorithm or one-gram algorithm based on the query size.
- Match the query with the token in the inverted index, compute corresponding joint relevance scores from features of Anchor Text, TF-IDF of Title & Body text and PageRank respectively.
- In two-gram algorithm, List of matched document is divided into two buckets with the prior one containing the two-gram tokens and the other one containing either of the query word.
- Ranking inside the prior bucket are based on both two-gram and one-gram relevance scores from features of Anchor Text, TF-IDF of Title & Body text. Assigned these two parts with different weights and join them together with Page Rank values of the document set to finalized the total score.
- Ranking of the second bucket is same as the one-gram algorithm.
- Return top 10 results with highest relevance scores.

3: NDCG computation

We use Discounted cumulative gain (DCG) to measure the ranking quality of our search engine, the formula is shown below:

$$DCG_i = rel_i + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

DCG score is normalized taking Google search result as the ground truth. The graded relevance scale of top 5 documents in the search result set ranges from 5 to 1.

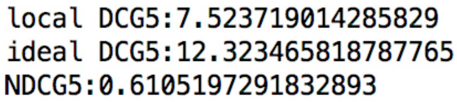
For example, when we search query “Crista Lopes” in Google, the top five results are :

http://www.ics.uci.edu/~lopes
http://www.ics.uci.edu/about/search/search_dept_in4matx.php
http://www.ics.uci.edu/~lopes/patents.html
http://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes
http://mondego.ics.uci.edu

When we search the same query in our search engine, the top five results are:

http://www.ics.uci.edu/~lopes
http://www.ics.uci.edu/~lopes/patents.html
http://www.ics.uci.edu/~lopes/aop/aop-pics.html
http://www.ics.uci.edu/~lopes/aop/aop.html
http://www.ics.uci.edu/~lopes/dv/dv.html

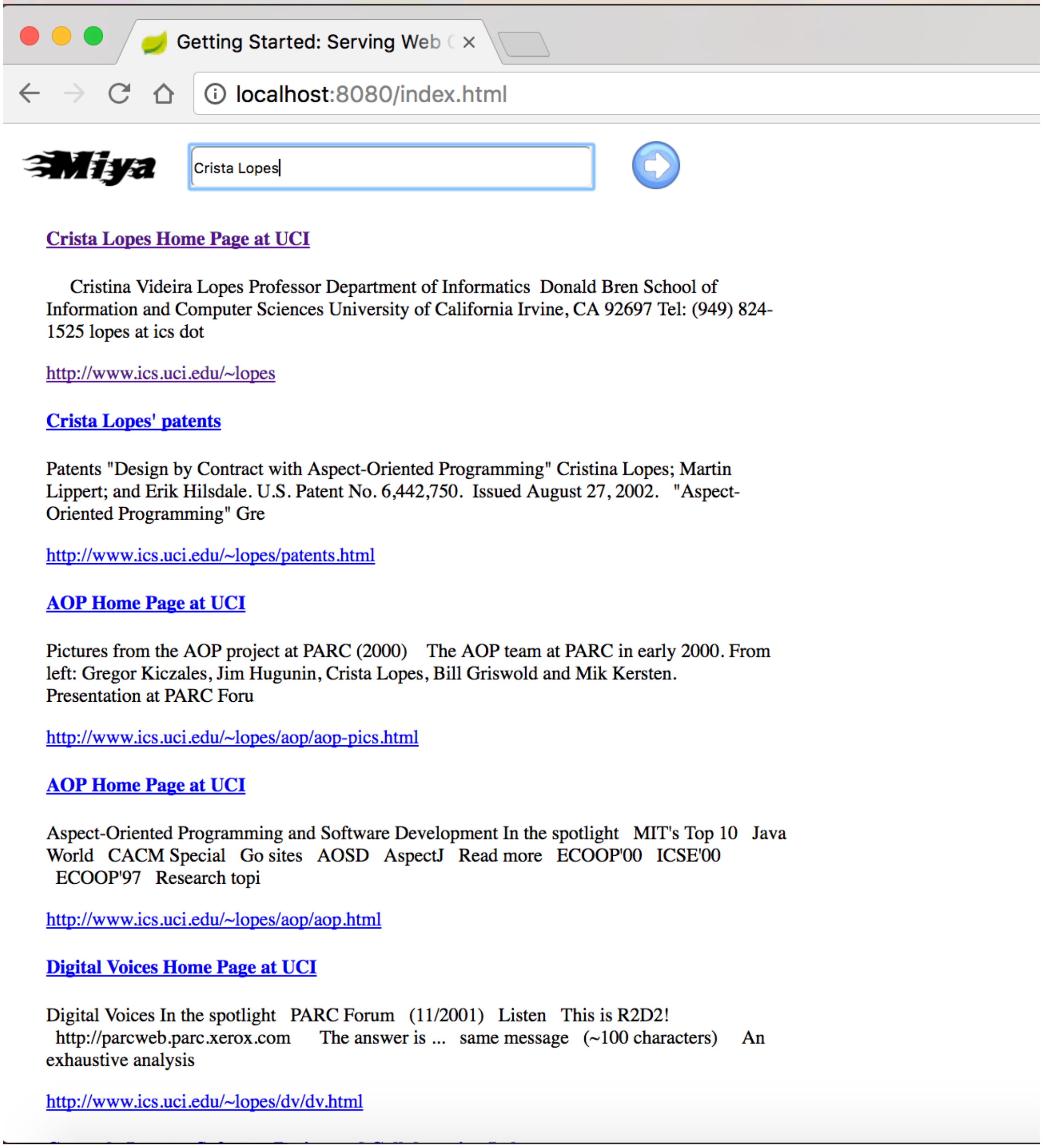
Then we The snapshot of NDCG result on query “Crista Lopes” is shown in picture below.



```
local DCG5:7.523719014285829
ideal DCG5:12.323465818787765
NDCG5:0.6105197291832893
```

4.Search Engine Result

The search result for query “Crista Lopes” from our search engine is shown below.



[Crista Lopes Home Page at UCI](http://www.ics.uci.edu/~lopes)

Cristina Videira Lopes Professor Department of Informatics Donald Bren School of Information and Computer Sciences University of California Irvine, CA 92697 Tel: (949) 824-1525 lopes at ics dot

http://www.ics.uci.edu/~lopes

[Crista Lopes' patents](http://www.ics.uci.edu/~lopes/patents.html)

Patents "Design by Contract with Aspect-Oriented Programming" Cristina Lopes; Martin Lippert; and Erik Hilsdale. U.S. Patent No. 6,442,750. Issued August 27, 2002. "Aspect-Oriented Programming" Gre

http://www.ics.uci.edu/~lopes/patents.html

[AOP Home Page at UCI](http://www.ics.uci.edu/~lopes/aop/aop-pics.html)

Pictures from the AOP project at PARC (2000) The AOP team at PARC in early 2000. From left: Gregor Kiczales, Jim Hugunin, Crista Lopes, Bill Griswold and Mik Kersten. Presentation at PARC Foru

http://www.ics.uci.edu/~lopes/aop/aop-pics.html

[AOP Home Page at UCI](http://www.ics.uci.edu/~lopes/aop/aop.html)

Aspect-Oriented Programming and Software Development In the spotlight MIT's Top 10 Java World CACM Special Go sites AOSD AspectJ Read more ECOOP'00 ICSE'00 ECOOP'97 Research topi

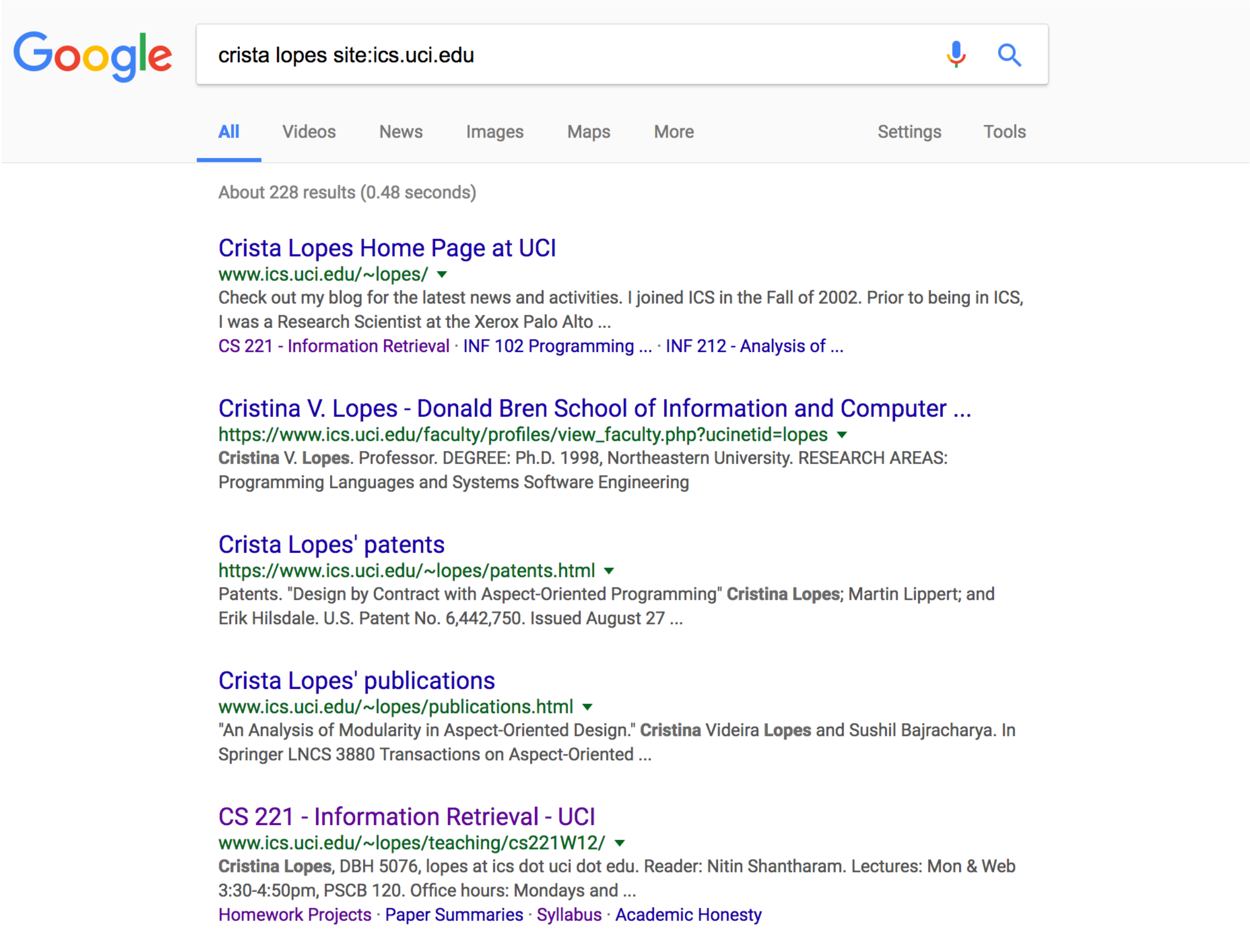
http://www.ics.uci.edu/~lopes/aop/aop.html

[Digital Voices Home Page at UCI](http://www.ics.uci.edu/~lopes/dv/dv.html)

Digital Voices In the spotlight PARC Forum (11/2001) Listen This is R2D2! http://parcweb.parc.xerox.com The answer is ... same message (~100 characters) An exhaustive analysis

http://www.ics.uci.edu/~lopes/dv/dv.html

The corresponding search result from google is shown below.



About 228 results (0.48 seconds)

Crista Lopes Home Page at UCI
www.ics.uci.edu/~lopes/
Check out my blog for the latest news and activities. I joined ICS in the Fall of 2002. Prior to being in ICS, I was a Research Scientist at the Xerox Palo Alto ...
CS 221 - Information Retrieval · INF 102 Programming ... · INF 212 - Analysis of ...

Cristina V. Lopes - Donald Bren School of Information and Computer ...
https://www.ics.uci.edu/faculty/profiles/view_faculty.php?ucinetid=lopes
Cristina V. **Lopes**. Professor. DEGREE: Ph.D. 1998, Northeastern University. RESEARCH AREAS: Programming Languages and Systems Software Engineering

Crista Lopes' patents
https://www.ics.uci.edu/~lopes/patents.html
Patents. "Design by Contract with Aspect-Oriented Programming" **Cristina Lopes**; Martin Lippert; and Erik Hilsdale. U.S. Patent No. 6,442,750. Issued August 27 ...

Crista Lopes' publications
www.ics.uci.edu/~lopes/publications.html
"An Analysis of Modularity in Aspect-Oriented Design." **Cristina** Videira **Lopes** and Sushil Bajracharya. In Springer LNCS 3880 Transactions on Aspect-Oriented ...

CS 221 - Information Retrieval - UCI
www.ics.uci.edu/~lopes/teaching/cs221W12/
Cristina Lopes, DBH 5076, lopes at ics dot uci dot edu. Reader: Nitin Shantharam. Lectures: Mon & Web 3:30-4:50pm, PSCB 120. Office hours: Mondays and ...
Homework Projects · Paper Summaries · Syllabus · Academic Honesty

5.Further Improvement

Based on all three steps of Project 3, we realized the search engine with relatively satisfying search results on given queries. For further improvement, the URL pool may be extended thus to cover the entire subdomain as needed. Also, to strengthen the PageRank factor by inserting more hub links can also help with precision. What's more, more file types should be supported, for instance, files in pfd or ppt play an important role in academic search.

In conclusion, we successfully implemented a complete search engine for the ICS domain. Specifically, we generated inverted index that maps words to documents, then constructed an interface for search queries and returned top 10 results with highest relevance score. Finally, optimized relevance algorithm is implemented and the ranking performance has been improved.