Bringing Undergraduate Students Closer to a Real-World Information Retrieval Setting: Methodology and Resources

Julián Urbano, Mónica Marrero, Diego Martín and Jorge Morato

http://julian-urbano.info
Twitter: @julian_urbano
Information Retrieval

• **Automatically** search documents
  ▫ **Effectively**
  ▫ **Efficiently**

• Expected between 2009 and 2020 [Gantz et al., 2010]
  ▫ 44 times more digital information
  ▫ 1.4 times more staff and investment

• IR is very **important for the world**
• IR is clearly important for Computer Science students
  ▫ Usually **not in the core** program of CS majors
    [Fernández-Luna et al., 2009]
Experiments in Computer Science

- Decisions in the CS industry
  - Based on mere experience
  - Bias towards or against some technologies

- CS professionals need background
  - Experimental methods
  - Critical analysis of experimental studies [IEEE/ACM, 2001]

- IR is a highly experimental discipline [Voorhees, 2002]
- Focus on evaluation of search engine effectiveness
  - Evaluation experiments not common in IR courses [Fernández-Luna et al., 2009]
Proposal: Teach IR through Experimentation

• Try to resemble a real-world IR scenario
  ▫ Relatively large amounts of information
  ▫ Specific characteristics and needs
  ▫ Evaluate what techniques are better

• Can we reliably adapt the methodologies of academic/industrial evaluation workshops?
  ▫ TREC, INEX, CLEF...
Current IR Courses

- Develop IR systems **extending** frameworks/tools
  - IR Base [Calado et al., 2007]
  - Alkaline, Greenstone, SpidersRUs [Chau et al., 2010]
  - Lucene, Lemur

- Students should **focus on the concepts** and **not struggle** with the tools [Madnani et al., 2008]
Current IR Courses (II)

- Develop IR systems **from scratch**
  - Build your search engine in 90 days [Chau et al., 2003]
  - History Places [Hendry, 2007]

- Only possible in **technical majors**
  - Software development
  - Algorithms
  - Data structures
  - Database management

- Much **more complicated** than using other tools
- Much **better to understand** the IR techniques and processes explained
Current IR Courses (III)

- Students **must** learn the **Scientific Method**
  - [IEEE/ACM Computing Curricula, 2001]

- Not only get the numbers [Markey et al., 1978]
  - Where do they come from?
  - What do they really mean?
  - Why they are calculated that way?

- Analyze the **reliability and validity**

- Not usually paid much attention in IR courses
  - [Fernández-Luna et al., 2009]
Current IR Courses (and IV)

• Modern CS industry is changing
  ▫ **Big data**

• Real conditions differ widely from class [Braught et al., 2004]
  ▫ Interesting from the experimental viewpoint

• Finding **free collections** to teach IR is hard
  ▫ **Copyright**
  ▫ Diversity

• Usually **stick** to the same collections **year after year**
  ▫ Try to make them a **product of the course** itself
Our IR Course

- Senior CS undergraduates
  - So far: elective, 1 group of ~30 students
  - From next year: mandatory, 2-3 groups of ~35 students

- Exercise: very simple Web crawler [Urbano et al., 2010]
- Search engine form scratch, in groups of 3-4 students
  - Module 1: basic indexing and retrieval
  - Module 2: query expansion
  - Module 3: Named Entity Recognition
- Skeleton and evaluation framework are provided
- External libraries allowed for certain parts
  - Stemmer, Wordnet, etc.

- Microsoft .net, C#, SQL Server Express
Evaluation in Early TREC Editions

Document collection, depending on the task, domain...
Evaluation in Early TREC Editions

Document collection, depending on the task, domain...

Relevance assessors, retired analysts
Evaluation in Early TREC Editions

Document collection, depending on the task, domain...

Candidate topics
Evaluation in Early TREC Editions

Document collection, depending on the task, domain...

Topic difficulty?
Evaluation in Early TREC Editions

Document collection, depending on the task, domain...

Organizers choose final ~50 topics
Evaluation in Early TREC Editions
Evaluation in Early TREC Editions
Evaluation in Early TREC Editions

Top 1000 results per topic
Evaluation in Early TREC Editions
Evaluation in Early TREC Editions

Top 100 results
Evaluation in Early TREC Editions

Top 100 results

Depth-100 pool

Depending on overlapping, pool sizes vary (usually ~1/3)
Evaluation in Early TREC Editions

Top 100 results

Depth-100 pool

Depending on overlapping, pool sizes vary (usually ~1/3)

What documents are relevant?
Evaluation in Early TREC Editions

Top 100 results

Depth-100 pool

Depending on overlapping, pool sizes vary (usually ~1/3)

What documents are relevant?

Relevance judgments

Organizers
Evaluation in Early TREC Editions

Top 100 results

Depth-100 pool

Depending on overlapping, pool sizes vary (usually ~1/3)

Relevance judgments

What documents are relevant?

Organizers
Changes for the Course

Document collection, can not be too large
Changes for the Course

• Different themes = many diverse documents
• Similar topics = fewer documents

• Start with a theme common to all topics

Document collection, can not be too large

Topics are restricted
Changes for the Course

How do we get **relevant** documents?

Write topics with a common theme
Changes for the Course

Write topics with a common theme

Try to answer the topic ourselves
Changes for the Course

Write topics with a common theme

Try to answer the topic ourselves

Too difficult? Too few documents?
Changes for the Course

Write topics with a common theme

Try to answer the topic ourselves

Too difficult? Too few documents?

Final topics

Complete document collection
Changes for the Course

• Relevance judging would be next

• **Before** students start developing their systems
  ▫ **Cheating** is very tempting: all (mine) relevant!

• **Reliable pools** without the student systems?

• Use other well-known systems: Lucene & Lemur
  ▫ These are the **pooling systems**
  ▫ Different techniques, similar to what we teach
Changes for the Course

Lucene  Lemur  Lucene  ...  Lemur
Changes for the Course

Some students would judge many more documents than others.

Lucene  Lemur  Lucene  Lemur

Depth-k pools lead to very different sizes.
Changes for the Course

• Instead of depth-k, pool until size-k
  ▫ **Size-100** in our case
  ▫ About 2 hours to judge, **possible to do in class**
Changes for the Course

• Pooling systems might leave relevant material out
  ▫ Include Google’s top 10 results to each pool
    • When choosing topics we knew there were some
• Students might do the judgments carelessly
  ▫ Crawl 2 noise topics with unrelated queries
  ▫ Add 10 noise documents to each pool and check

• Now keep pooling documents until size-100
  ▫ The union of the pools form the biased collection
    • This is the one we gave students
2010 Test Collection

• 32 students, 5 faculty

• Theme: Computing
  ▫ 20 topics (plus 2 for noise)
    • 17 judged by two people
  ▫ 9,769 documents (735MB) in complete collection
  ▫ 1,967 documents (161MB) in biased collection

• Pool sizes between 100 and 105
  ▫ **Average pool depth 27.5**, from 15 to 58
Evaluation

• With results and judgments, rank student systems

• Compute mean NDCG scores [Järvelin et al., 2002]
  ▫ Are relevant documents properly ranked?
  ▫ Ranges between 0 (useless) to 1 (perfect)

• The group with the best system is rewarded with an extra point (over 10) in the final grade
  ▫ About half the students really motivated for this
Is This Evaluation Method Reliable?

• These evaluations have problems [Voorhees, 2002]
  ▫ **Inconsistency of judgments** [Voorhees, 2000]
    • People have different notions of «relevance»
    • Same document judged differently
  ▫ **Incompleteness of judgments** [Zobel, 1998]
    • Documents not in the pool are considered not relevant
    • What if some of them were actually relevant?

• TREC ad hoc was quite reliable
  ▫ Many topics
  ▫ Great man-power
Assessor Agreement

• 17 topics were judged by 2 people
  ▪ Mean Cohen’s Kappa = 0.417
    • From 0.096 to 0.735
  ▪ Mean Precision-Recall = 0.657 - 0.638
    • From 0.111 to 1

• For TREC, observed 0.65 precision at 0.65 recall
  [Voorhees, 2000]

• Only 1 noise document was judged relevant
Inconsistency & System Performance

![Graph showing mean NDCG@100 over 2,000 trels]

- Mean over 2,000 trels
Inconsistency & System Performance (and II)

• Take 2,000 random samples of assessors
  ▫ Each would have been possible if using 1 assessor
  ▫ NDCG differences between 0.075 and 0.146

• 1.5 times larger than in TREC [Voorhees, 2000]
  ▫ Unexperienced assessors
  ▫ 3-point relevance scale
  ▫ Larger values to begin with! (x2)
    • Percentage differences are lower
Inconsistency & System Ranking

• Absolute numbers do not tell much, **rankings** do
  ▫ It is highly dependent on the collection

• Compute correlations with 2,000 random samples
  ▫ Mean Kendall’s tau = **0.926**
    • From 0.811 to 1

• For TREC, observed **0.938** [Voorhees, 2000]

• **No swap was significant** (Wilcoxon, \(\alpha=0.05\))
Incompleteness & System Performance

Mean difference over 2,000 trels
Estimated difference

Increment in Mean NDCG@100 (%) vs. Pool size
Incompleteness & System Performance (and II)

- Start with size-20 pools and keep adding 5 more
  - 20 to 25: NDCG variations between 14% and 37%
  - 95 to 100: between 0.4% and 1.6%, mean 1%

- In TREC editions, between 0.5% and 3.5% [Zobel, 1998]
  - Even some observations of 19%!

- We achieve these levels for sizes 60-65
Incompleteness & Effort

- **Extrapolate** to larger pools
  - If differences decrease a lot, judge a little more
  - If not, judge fewer documents but more topics

- **135** documents for mean NDCG differences <0.5%
  - Not really worth the +35% effort
Student Perception

- Brings more work both for faculty and students

- More **technical problems** than previous years
  - Survey scores dropped from 3.27 to 2.82 over 5
  - Expected, they had to work considerably more
  - **Possible gaps in (our) CS program?**

- **Same satisfaction scores as previous years**
  - Very appealing for us
  - This was our first year, lots of logistics problems
Conclusions & Future Work

• IR deserves more attention in CS education
• Laboratory experiments deserve it too

• **Focus on the experimental nature of IR**
  ▫ But give a wider perspective useful beyond IR

• Much more **interesting and instructive**

• Adapt well-known methodologies in industry/academia to our limited resources
Conclusions & Future Work (II)

• Improve the methodology
  ▫ Explore quality control techniques in crowdsourcing

• Integrate with other courses
  ▫ Computer Science
  ▫ Library Science

• Make everything public, free
  ▫ Tets collections and reports
  ▫ [http://ir.kr.inf.uc3m.es/eirex/](http://ir.kr.inf.uc3m.es/eirex/)
Conclusions & Future Work (and III)

• We call all this EIREX
  ▫ Information Retrieval Education through Experimentation

• We would like others to participate
  ▫ Use it in your class
    • Collections
    • The methodology
    • Tools
I WANT YOU FOR EIREX
2011 Semester

• Only 19 students

• The 2010 collection used for testing
  ▫ Every year there will be more variety to choose

• 2011 collection: Crowdsourcing
  ▫ 23 topics, better described
    • Created by the students
    • Judgments by 1 student
      ▫ 13,245 documents (952MB) in complete collection
      ▫ 2,088 documents (96MB) in biased collection

• Overview report coming soon!
Can we **reliably** adapt the large-scale methodologies of academic/industrial IR evaluation workshops?

**YES (we can)**