Grading Search Engines like Google and Bing

Ashwin Satyanarayana

Department of Computer Systems Technology

Note: This was work done when I was at Microsoft

School Assignment Satisfaction Index (SASI)

- Judge see the search engine results pages (**SERP**)
- They do not see landing pages
  - But do judge captions
- Judging is in SERP context
- Cheap, one-off experimentation
- Some abstraction
- Evaluation by block, element, quality

Microsoft, Inc. © 2010 — 1

Judge Based Evaluation

- Label relevance of docs/SERP components to:
  - Acquire training, and (reusable) validation & measurement data
  - Enable relevance metrics
- Measure to:
  - Identify weaknesses → opportunities to improve
  - Determine which experience to ship
  - Compare our work to our competitors
- Provide a framework for communication and discussion of predicted user satisfaction

Microsoft, Inc. © 2010

Whole Page Relevance

- *Whole Page Relevance* defines how well search engine result page components and the corresponding holistic attributes of the result page presentation respond to searchers' information needs

Microsoft, Inc. © 2010 — 3

School assignment grading activity

School assignments
- Grades
  - A+
  - 14/20
  - Pass
- Characterize response by
  - Answered question correctly
  - Displayed comprehension
  - Wrote fluently, grammatically
  - Covered topic aspects
  - Good structure
  - …

Microsoft, Inc. © 2010 — 4
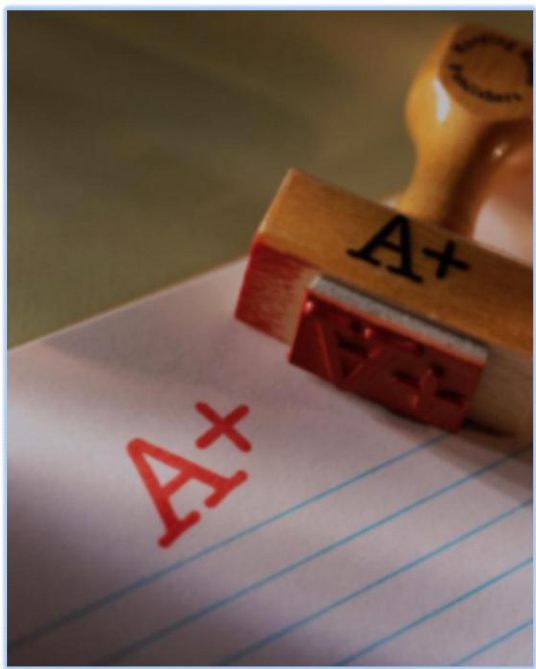
Result page grading aspects

Content
- Block or component-level relevance
  - E.g. Left-rail, Related searches
- Element-level relevance
  - E.g. Web doc result 5, Top mainline advertisement
- Sub-element level relevance
  - E.g. Caption, URL
- Graded relevance per item

Holistic
- Freshness
- Diversity
- Authority
- Cross-element topic consistency
- Redundancy
- …

Framework for evaluation

Framework
- Ability to pull apart SERP into components and sub-components
- Element "highlighting"
- Rating facility in SERP context for:
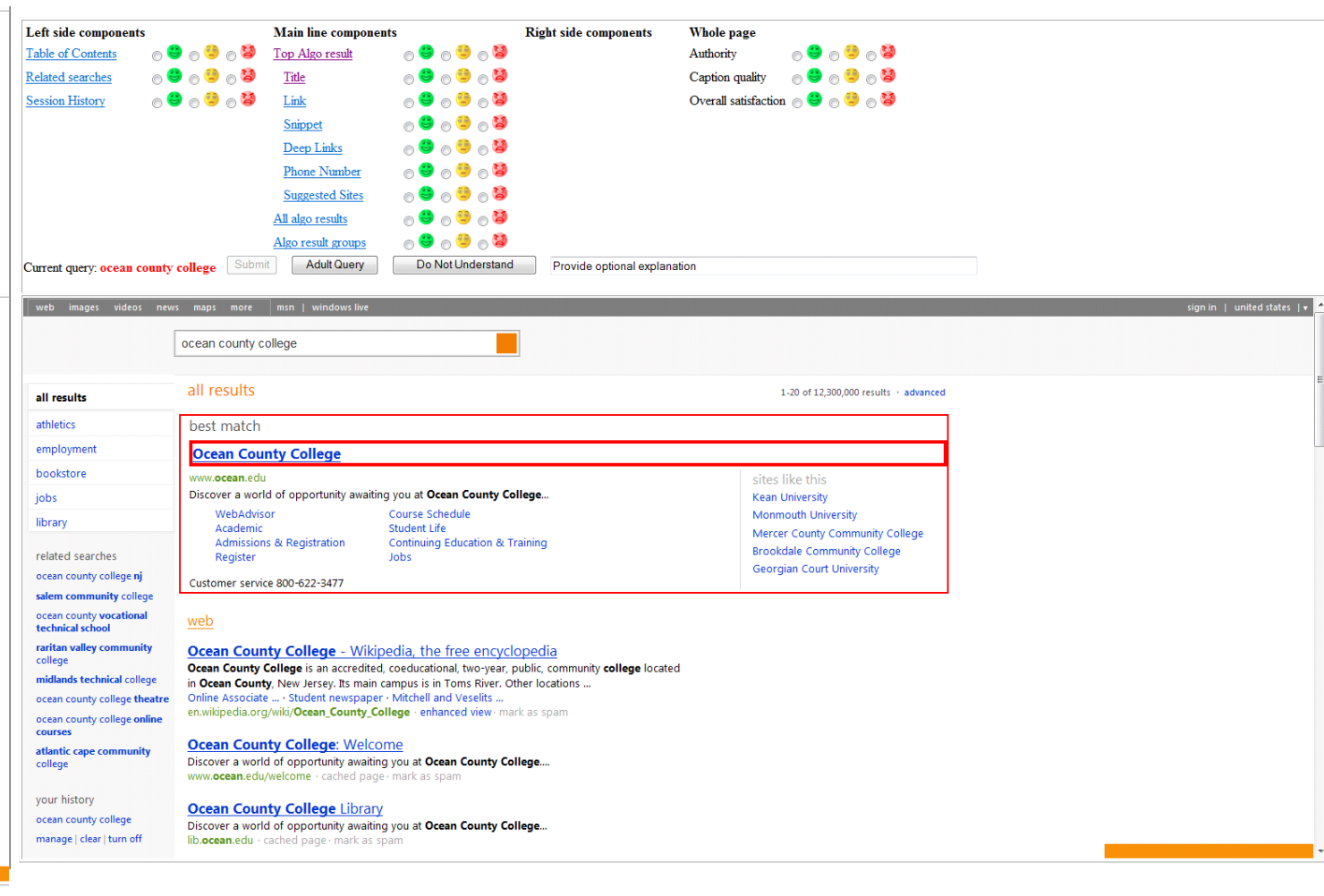 - Element contents
  - Holistic attributes

Method
1. Specify experiment
   a) Identify a set of queries
   b) **Specify blocks and components**
   c) **Specify holistic attributes**
2. **Capture SERPs, identify layout**
3. Allocate queries to judges
4. **Judge interaction**
   a) **Views query and SERP**
   b) **Rates blocks and components**
   c) **Rates holistic attributes**

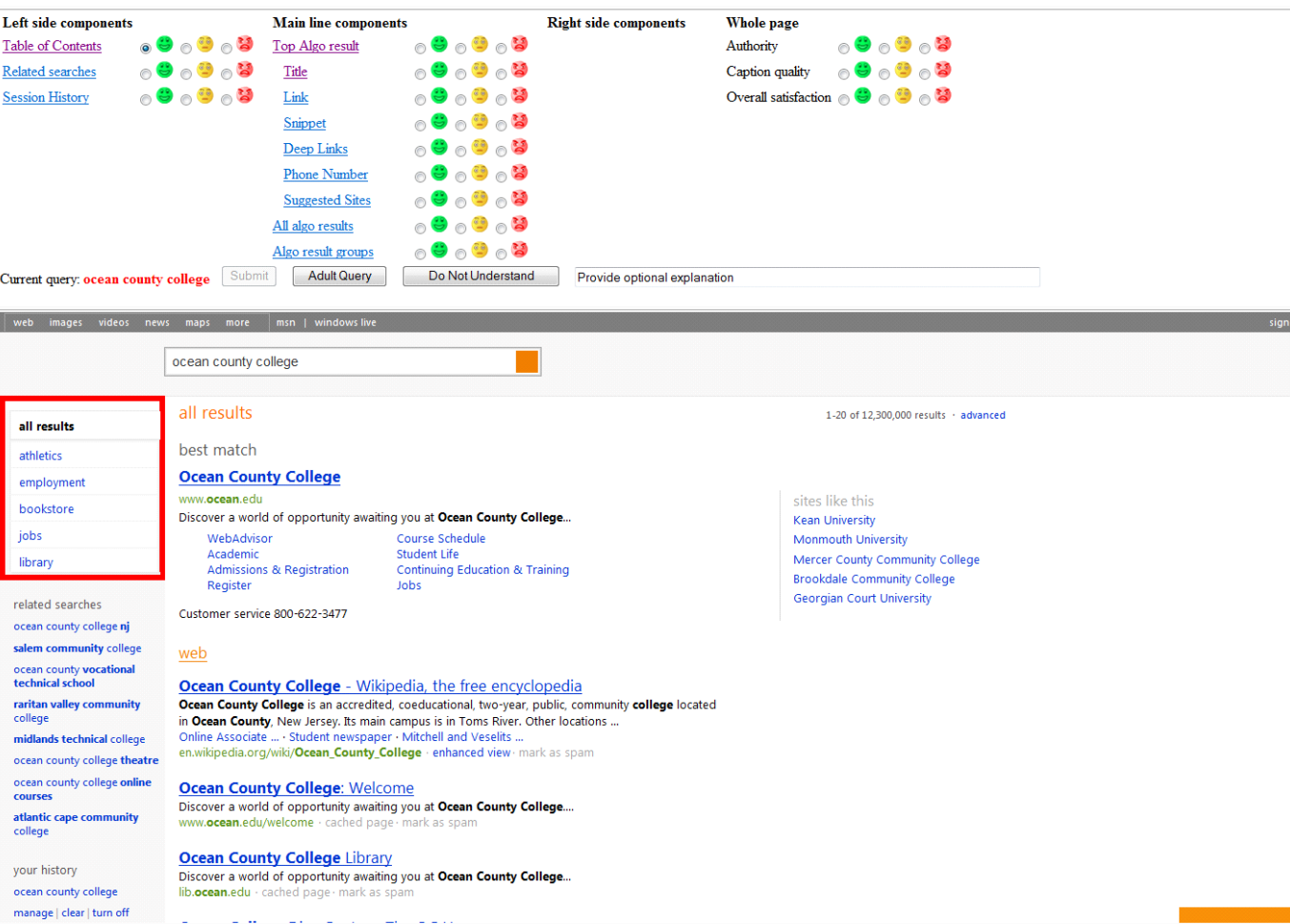Microsoft, Inc. © 2010 — 6
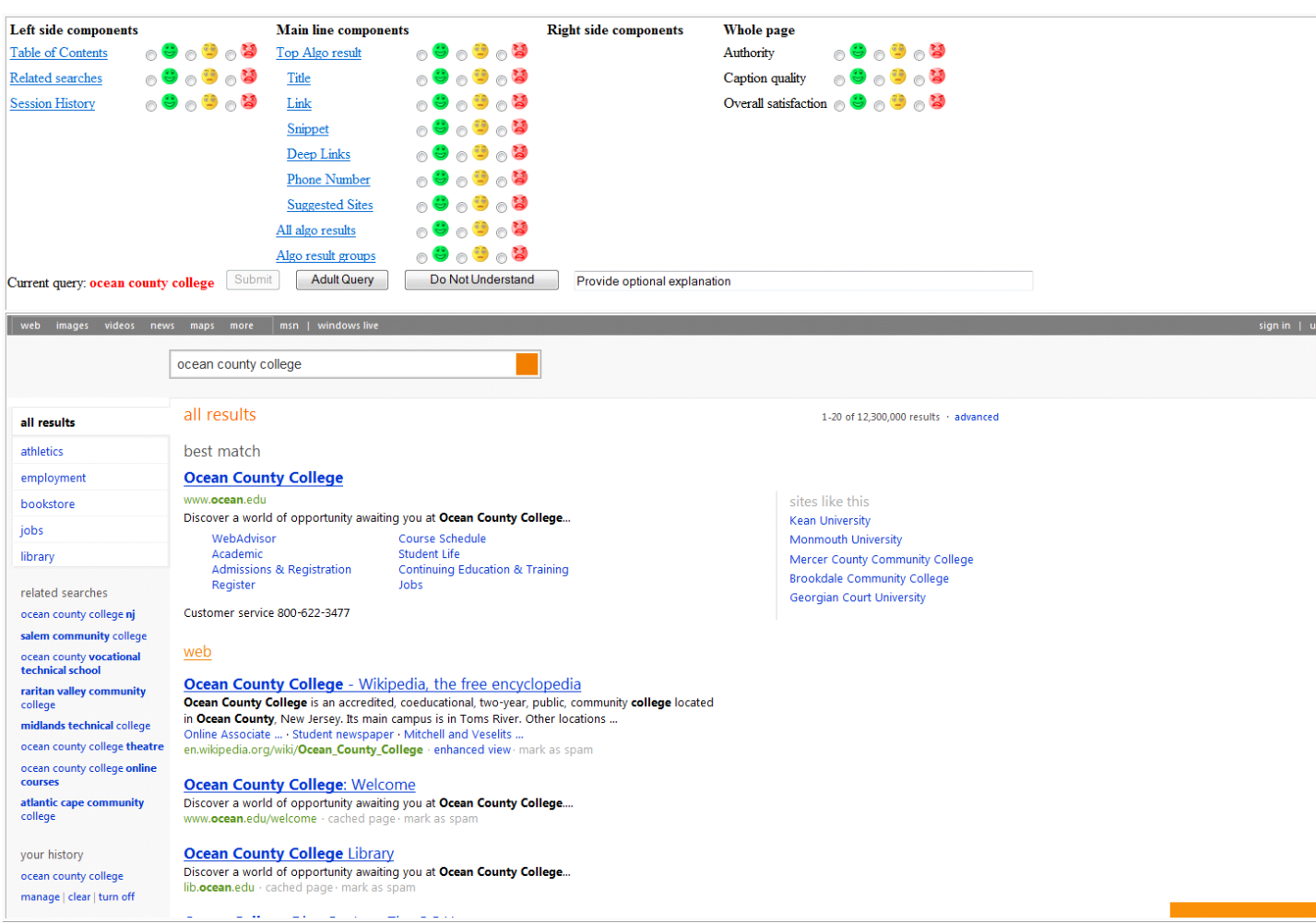
Example SASI UI - query is [ocean county college]
Rating components auto-configure to match SERP structure
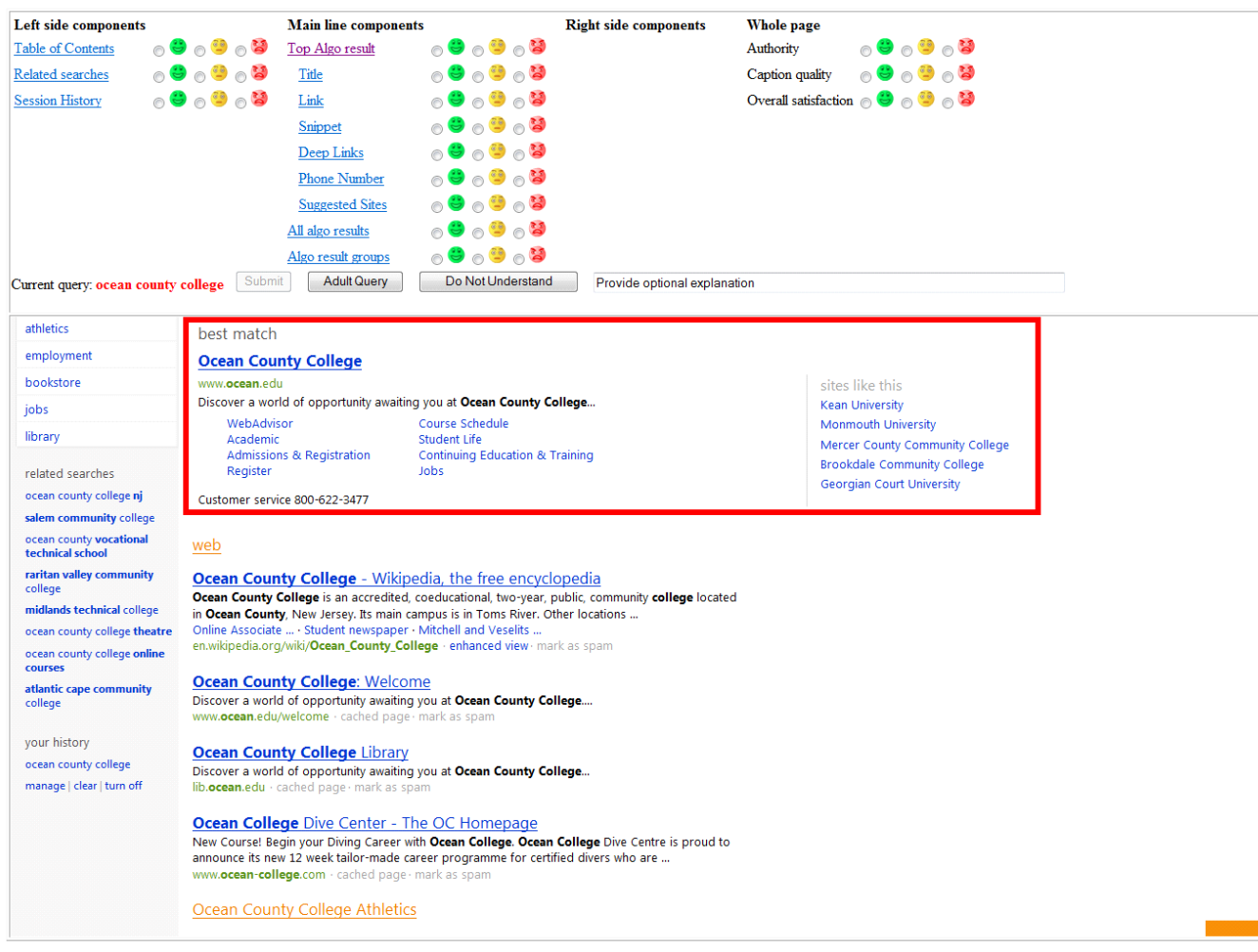
Clicking on a component link highlights and scrolls the page to the start of it

Microsoft, Inc. © 2010

Sub-element breakdown is possible – highlighting the result caption title

Microsoft, Inc. © 2010 — 9

The judge rates each (sub)component on a 3-point color-coded emoticon scale

Microsoft, Inc. © 2010 — 10

Content and holistic quality

Table 1. Avg. scores for components and holistic qualities.
Statistical significance levels are noted in sub/superscript ($\dagger$ < 0.05; $\ddagger$ < 0.01). The preceding letter denotes the comparator engine with which the difference is significant.

Comp. or Quality	A	B	C	Range
Query suggestions	1.88		1.71	0.21
Right-side ads	1.76	1.61		0.16
Diversity		1.75	1.64	0.12
Caption quality		1.65	1.59	0.08
Top result		1.91	1.87	0.06
Result 5	1.80		1.80	0.06
Result 4	1.81		1.80	0.06
Result 2	1.86		1.83	0.06
Overall satisfaction			1.80	0.06
Result 3		1.85	1.83	0.04

Microsoft, Inc. © 2010

Papers Published based on this work:

- **Evaluating Whole-Page Relevance**
  33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010), Pages: 767-768

- **Evaluating Search Systems Using Result Page Context (***Best Paper Finalist***)**
  3rd Symposium on Information Interaction in Context (IIiX 2010), Pages: 105-114