CloudSearch and the Democratization of Information Retrieval

Daniel E. Rose
A9.com
danrose@a9.com
What Does A9 Do?

Product Search

Visual Search

Advertising Technology

Community Q&A

15 August 2012

D. Rose, CloudSearch, SIGIR 2012
... and CloudSearch

A new hosted search service offered by AWS
Democratizing Information Retrieval: A Brief History
Democratizing Information Retrieval

• Giving more **users** access to search tools, and making those tools easier to use and more powerful

• Giving more **content owners** (businesses, organizations, research teams, government offices, etc.) the ability to be **search providers**.
1970s: Online Metered Search Services

• Examples: Dialog, ORBIT, Lexis/Nexis, Westlaw, BRS

• Cost and requirements [users]:
  – Installation and rental of dedicated terminal
  – Usage cost per hour (e.g. $50)
  – Cost per page printed, etc.

• Content available: Research corpora (e.g. journal articles), news stories, court cases.

• Improved access for:
  – Users (researchers, lawyers, etc.)
1970s: Online Metered Search Services

• Typical Query:

  assum! /5 risk /p ic* snow***
  snowfall /s slip! fell fall***

• Results: Often the first screen of the first retrieved document.

• Restrictions encouraged batch-style search
1980s: Enterprise Search Products

- Examples: Verity Topic, Personal Library Systems, Fulcrum SearchServer, Excalibur RetrievalWare.

- Cost and requirements:
  - $10-100K per year license fee, also per seat
  - Beefy hardware to install it on

- Improved access for:
  - Content owners (usually large businesses).
1990s: Web Search

• Examples: WebCrawler, Lycos, Infoseek, AltaVista, Excite, Inktomi, Yandex, Google, AllTheWeb, Teoma.

• Cost and requirements:
  – For users: Free, web browser.
  – For search providers: web server, high-speed service

• Improved access for:
  – Users
  – Content owners (as long as your data was HTML, and you put it on your website, and search engine chose to crawl it)
2000s: Open Source Search

• Examples: Lucene/Solr, Indri

• Cost and requirements [providers]:
  – No cost for software
  – Need hardware to run it on

• Improved Access for:
  – Content owners (with resources and expertise.)
2010s: CloudSearch

• Put your content in the cloud and make it searchable
• You decide what content gets searched and who can see it
• Self-service
• Improves access for:
  – Content owners
An Introduction to CloudSearch
What Is Amazon CloudSearch?

• A hosted web search service developed by A9
• Powered by the same search engine used by Amazon.com and other retailers.
• Designed from ground up to support:
  – semistructured data
  – faceted metadata search
  – numeric range searches
  – memory-resident indexes

15 August 2012
D. Rose, CloudSearch, SIGIR 2012
CloudSearch Uses AWS Services

- Elastic Compute Cloud (EC2) for computation
- Simple Storage Service (S3) for storage
- Elastic Map Reduce (EMR) for index construction
- Simple Work Flow (SWF) for coordinating customer actions
- Elastic Load Balancing (ELB) for routing traffic
Setting Up the Data

CloudSearch can help you define your index fields automatically using a **configuration wizard** or you can manually edit the configuration below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Type</th>
<th>Search</th>
<th>Facet</th>
<th>Result</th>
<th>Default Value</th>
<th>Source</th>
<th>Remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td>Active</td>
<td>text</td>
<td>✓</td>
<td>⬜</td>
<td>✓</td>
<td></td>
<td>[add]</td>
<td></td>
</tr>
<tr>
<td>director</td>
<td>Active</td>
<td>text</td>
<td>✓</td>
<td>⬜</td>
<td>✓</td>
<td></td>
<td>[add]</td>
<td></td>
</tr>
<tr>
<td>genre</td>
<td>Active</td>
<td>literal</td>
<td>✓</td>
<td>✓</td>
<td>⬜</td>
<td></td>
<td>[add]</td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Active</td>
<td>text</td>
<td>✓</td>
<td>⬜</td>
<td>✓</td>
<td></td>
<td>[add]</td>
<td></td>
</tr>
<tr>
<td>year</td>
<td>Active</td>
<td>uint</td>
<td>⬜</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>[add]</td>
<td></td>
</tr>
</tbody>
</table>

[Add Index Field]
Indexing Documents
(additions, updates, deletions)
# Testing Queries from Dashboard

## Run a Test Search in Domain drtest1

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Year</th>
<th>Director</th>
<th>Actors</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>tt0073195</em></td>
<td>1975</td>
<td>Spielberg, Steven</td>
<td>Anastasio, Chris; Backlinie, Susan; Chambers, Robert; Cheshire, Denise; Courtney, Fritz; Dreyfuss, Richard; Garr, Teri; Guffey, Steve</td>
<td>Drama (14)</td>
</tr>
<tr>
<td>2</td>
<td><em>tt0075860</em></td>
<td>1977</td>
<td>Spielberg, Steven</td>
<td>Balaban, Bob; Blossom, Roberts; Dillon, Melinda; Dodds, Philip; Dreyfuss, Richard; Garr, Teri; Guffey, Steve</td>
<td>Drama (14)</td>
</tr>
<tr>
<td>3</td>
<td><em>tt0078723</em></td>
<td>1979</td>
<td>Spielberg, Steven</td>
<td>Aykroyd, Dan; Beatty, Ned; Belushi, John; Caan, James; Cheshire, Denise; Gary, Lorraine; Hamilton, Michael</td>
<td>Drama (14)</td>
</tr>
<tr>
<td>4</td>
<td><em>tt0082971</em></td>
<td>1981</td>
<td>Spielberg, Steven</td>
<td>Allen, Karen; Diamond, Peter; Elliott, Denholm; Fielder, Harry; Ford, Harrison; Freeman, Paul; Gifford, Steve</td>
<td>Drama (14)</td>
</tr>
<tr>
<td>5</td>
<td><em>tt0083886</em></td>
<td>1983</td>
<td>Spielberg, Steven</td>
<td>Allen, Karen</td>
<td>Drama (14)</td>
</tr>
</tbody>
</table>
Testing Queries from Web Browser

<results>
  <rank>year</rank>
  <match-expr>(label director;spielberg')</match-expr>
  <hits found='27' start='0'>
    <hit id='10073193'>
      <cd name='director'>Spielberg, Steven</cd>
      <cd name='title'>Jaws</cd>
      <cd name='year'>1975</cd>
    </hit>
    <hit id='10075860'>
      <cd name='director'>Spielberg, Steven</cd>
      <cd name='title'>Close Encounters of the Third Kind</cd>
      <cd name='year'>1977</cd>
    </hit>
    <hit id='1008333'>
      <cd name='director'>J.C. EE</cd>
      <cd name='title'>Indiana Jones</cd>
      <cd name='year'>1989</cd>
    </hit>
  </hits>
  <facets>
    <facet name='genre'>
      <constraint value='Drama' count='14'/>
      <constraint value='Adventure' count='12'/>
      <constraint value='Action' count='11'/>
      <constraint value='Sci-Fi' count='8'/>
      <constraint value='Thriller' count='5'/>
      <constraint value='War' count='5'/>
      <constraint value='Family' count='4'/>
      <constraint value='History' count='4'/>
      <constraint value='Biography' count='3'/>
      <constraint value='Comedy' count='3'/>
      <constraint value='Fantasy' count='3'/>
      <constraint value='Mystery' count='3'/>
      <constraint value='Crime' count='2'/>
      <constraint value='Romance' count='2'/>
      <constraint value='Animation' count='1'/>
    </facet>
  </facets>
</results>
Search API

q = close+encounters

bq = (and (or director:'spielberg'
director:'lucas')
year:1975..1980)

rank = −year,title

return-fields: director,title,year

facet = genre
CloudSearch Relevance Ranking

• Configurable ranking functions
• Can combine \( tf \times idf \)-style text matching score with query-independent ranking features.
• Rank Expressions:

\[
0.4 \times \log_2\left(\frac{\text{time()}}{31536000000 - \text{year}}\right) + 0.6 \times \text{text_relevance}
\]
Elastic Scaling
To Recap Benefits:

• Easy to make any semi-structured data searchable
• Easy to set up and configure
• No hardware or software management
• Scalable and elastic

Anyone can be a search provider.
Implications for Search User Experience
Progress in Search User Experience 1972-1995

- From highly structured boolean queries to unstructured text
- From binary matching to relevance-ranking
- From batch-like to interactive
- From command line to GUI
- From monospace 80 x 24 text to rich presentation
Web Search in 1995

Hierarchical Fuzzy Logic Control of a Double Inverted Pendulum
Hierarchical Fuzzy Logic Control of a Double Inverted Pendulum. Reza Langan and Shubing Lei Center for Fuzzy Logic and Intelligent Systems Research...
http://celebria.tamu.edu/project/diplomadip.htm; size 8K - 21 Nov 95

Project Management with Fuzzy Logic
Edit a Project. Please enter details of the project. For more information, please select help. Project ID: Project Name: Priority: Start Date: Duration:...
http://www.ca.bris.ac.uk/~wong/Proj/approj.html; size 1K - 13 Oct 95

Project Management with Fuzzy Logic
Create a New Project. Please enter details of the project. For more information, please select help. Project ID: Project Name: Priority: Start Date:...
http://www.ca.bris.ac.uk/~wong/Proj/newproj.html; size 1K - 28 Sep 95

Project Management with Fuzzy Logic
Help. This program is design to give an accurate estimation of project risk analysis before and during the different phases of a project. It also allows...
http://www.ca.bris.ac.uk/~wong/Proj/help.html; size 1K - 13 Oct 95

CENTRAL NOTICE - Fuzzy Logic
Apologies to fuzzy logicians. In a fuzzy world, you might only "partially" know stuff. So there would be much information that is neither...
http://www.notice.com/whry/fuzzy.html; size 1K - 1 Apr 96

Archimedes Fuzzy Logic Development Tool
Archimedes Fuzzy Logic Development Tool. This product introduces many new functions for Fuzzy development. A graphical CAD surface to design the system...
http://www.merimedia.com/pvdi/archimedes/mdl/archfuzzy.htm; size 1K - 19 Jul 95

Machine Learning Games, Neural Networks, and Fuzzy Logic
Hardware & Software Capabilities

<table>
<thead>
<tr>
<th></th>
<th>1994</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>120 MHz Intel Pentium</td>
</tr>
<tr>
<td>Memory</td>
<td>512 KB</td>
</tr>
<tr>
<td>OS</td>
<td>Windows 3.1</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>28.8 kbps</td>
</tr>
<tr>
<td>Cost</td>
<td>$2000</td>
</tr>
</tbody>
</table>

- Search engine designers passed the absolute minimum needed to decide whether to click: title, URL, small excerpt.
- Being able to point and click on remote content was a big deal!
## Hardware & Software Capabilities

<table>
<thead>
<tr>
<th>1994</th>
<th>2012</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>120 MHz Intel Pentium</td>
<td>2.5 GHz Intel Core i5</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>1 MB</td>
<td>4 GB</td>
</tr>
<tr>
<td><strong>OS</strong></td>
<td>Windows 3.1</td>
<td>Windows 7, Mac OS X</td>
</tr>
<tr>
<td><strong>Bandwidth</strong></td>
<td>28.8 kbps</td>
<td>5.8 Mbps</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>$2000</td>
<td>$500-1000</td>
</tr>
</tbody>
</table>

What have search engine UX designers done with all that additional power?
Web Search in 2012

![Google Search results for "fuzzy logic"](image)

- Fuzzy logic - Wikipedia, the free encyclopedia
- New York, NY
- Change location
- Show search tools
- FAQ: Fuzzy Logic and Fuzzy Expert Systems 1/1 [Monthly posting]
- piato.stanford.edu/entries/logic-fuzzy/
- Fuzzy Logic (Stanford Encyclopedia of Philosophy)
- videolectures.net
- More videos for fuzzy logic »
Why Lack of Progress?

• Most users’ first (and sometimes only) experience with search is with web search
• Search dominated by a few players
“What makes a good search engine user experience?”

- Results as relevant as possible
- Short delay between query and results
- Clean and uncluttered presentation
- Gives user a feeling of direct engagement
- Allows seamless transition between search and browsing
- Fun to use
- Rewards user for giving more information
- Interaction appropriate for type of task
- Limit visual noise / optimize data-ink ratio
- Minimizes scrolling

Optimizing for Other Properties
What does this have to do with CloudSearch?

• Unprecedented opportunity to build new search applications
• We’re not constrained by how web search works.

Not all search is web search.
Simple Illustrations with CloudSearch

Search the World Factbook with CloudSearch

- **Country**: Albania
- **Area (sq km)**: 28,748
- **Population**: 3,019,779
- **Life Expectancy (years)**: 77
- **Capital**: Tirana

Search the CIA World Factbook with CloudSearch

- **Country**: Aruba
- **Area (sq km)**: 186
- **Population**: 104,022
- **Life Expectancy (years)**: 74
- **Capital**: Oranjestad

The Bosnian Serbs were part of the Austro-Hungarian Empire until the latter's dissolution at the end of World War I in 1918. The Serbs were joined by Croats and Slovenes in forming a new multinational state, which was known as the Kingdom of Serbs, Croats, and Slovenes. In 1929, it became a republic of the newly formed Yugoslavia, which, though Communist, enjoyed self-rule from Moscow's rule. Dissatisfied with the exercise of power by the minority Serbs, the Slovenes succeeded in establishing their independence in 1991 after a short 10-day war and an increase in Western European, a strong economy, and a stable democracy. They associated themselves with the European Union and became a modern state. Slovenia acceded to both NATO and the EU in the spring of 2004.

The Bikini atoll is the site of the world's first atomic bomb tests conducted by the United States in 1946. The tests were conducted as part of Operation Crossroads, a program intended to test the effects of nuclear weapons on naval ships. The Bikini atoll, which is located in the Marshall Islands, was selected as the site of the tests because of its remote location and the belief that the radioactive fallout would be dispersed over a large area. The tests were conducted with two nuclear weapons: one, a 17-kiloton device called Able, was detonated on July 1, 1946, and the other, a 23-kiloton device called Baker, was detonated on July 25, 1946. The tests resulted in significant radioactive fallout, which contamined the atoll for many years. The Bikini atoll was designated a UNESCO World Heritage Site in 1986 to protect the environment and the cultural heritage of the islanders of the Marshall Islands.
A Typical Search Interface

Search the World Factbook with CloudSearch

Query: george w. bush

American Samoa
Settled as early as 1000 B.C., Samoa was "discovered" by European explorers in the 18th century. International rivalries in the latter half of the 19th century were settled by an 1899 treaty in which Germany and the US divided the Samoan archipelago. The ...

Guam
Guam was ceded to the US by Spain in 1898. Captured by the Japanese in 1941, it was retaken by the US three years later. The military installation on the island is one of the most strategically important US bases in the Pacific....

Northern Mariana Islands
Under US administration as part of the UN Trust Territory of the Pacific, the people of the Northern Mariana Islands decided in the 1970s not to seek independence but instead to forge closer links with the US. Negotiations for territorial status began in ...

Puerto Rico
Populated for centuries by aboriginal peoples, the island was claimed by the Spanish Crown in 1493 following COLUMBUS' second voyage to the Americas. In 1898, after 400 years of colonial rule that saw the indigenous population nearly exterminated and Afri...

United States
Britain's American colonies broke with the mother country in 1776 and were recognized as the new nation of the United States of America following the Treaty of Paris in 1783. During the 19th and 20th centuries, 37 new states were added to the original 13 ...

Virgin Islands
During the 17th century, the archipelago was divided into two territorial units, one English and the other Danish. Sugarcane, produced by slave labor, drove the islands' economy during the 18th and early 19th centuries. In 1917, the US purchased the Danis...
Results Can Be Interactive (and Contain Lots of Information)

Search the World Factbook with CloudSearch

<table>
<thead>
<tr>
<th>Country</th>
<th>Area (sq km)</th>
<th>Population</th>
<th>Life Expectancy</th>
<th>Capital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania</td>
<td>27398</td>
<td>3619778</td>
<td>77</td>
<td>Tirana (Tirane)</td>
</tr>
<tr>
<td>Romania</td>
<td>230340</td>
<td>22246862</td>
<td>72</td>
<td>Bucharest</td>
</tr>
<tr>
<td>Latvia</td>
<td>63589</td>
<td>2245423</td>
<td>71</td>
<td>Riga</td>
</tr>
<tr>
<td>Slovenia</td>
<td>20151</td>
<td>2007711</td>
<td>76</td>
<td>Ljubljana</td>
</tr>
<tr>
<td>Croatia</td>
<td>56414</td>
<td>4481543</td>
<td>75</td>
<td>Zagreb</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>110550</td>
<td>7262675</td>
<td>72</td>
<td>Sofia</td>
</tr>
<tr>
<td>Kosovo</td>
<td>10887</td>
<td>2126708</td>
<td>0</td>
<td>Pristina (Prishtine)</td>
</tr>
<tr>
<td>Hungary</td>
<td>92340</td>
<td>9930915</td>
<td>73</td>
<td>Budapest</td>
</tr>
<tr>
<td>Serbia</td>
<td>77474</td>
<td>10159046</td>
<td>75</td>
<td>Belgrade (Beograd)</td>
</tr>
</tbody>
</table>

The Slovene lands were part of the Austro-Hungarian Empire until the latter's dissolution at the end of World War I. In 1918, the Slovenes joined the Serbs and Croats in forming a new multinational state, which was named Yugoslavia in 1929. After World War II, Slovenia became a republic of the renewed Yugoslavia, which though Communist, distanced itself from Moscow's rule. Dissatisfied with the exercise of power by the majority Serbs, the Slovenes succeeded in establishing their independence in 1991 after a short 10-day war. Historical ties to Western Europe, a strong economy, and a stable democracy have assisted in Slovenia's transformation to a modern state. Slovenia acceded to both NATO and the EU in the spring of 2004.
Different Interface Controls in Different Situations

Search the World Factbook with CloudSearch

Query: sugar

<table>
<thead>
<tr>
<th>Country</th>
<th>Area (sq km)</th>
<th>Population</th>
<th>Life Expectancy</th>
<th>Capital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antigua and Barbuda</td>
<td>442</td>
<td>84522</td>
<td>74</td>
<td>Saint John's</td>
</tr>
<tr>
<td>Barbados</td>
<td>431</td>
<td>281968</td>
<td>73</td>
<td>Bridgetown</td>
</tr>
<tr>
<td>Belize</td>
<td>22806</td>
<td>301270</td>
<td>68</td>
<td>Belmopan</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>50660</td>
<td>4195914</td>
<td>77</td>
<td>San Jose</td>
</tr>
<tr>
<td>Cuba</td>
<td>110860</td>
<td>11423952</td>
<td>77</td>
<td>Havana</td>
</tr>
<tr>
<td>Dominican Republic</td>
<td>48380</td>
<td>9507133</td>
<td>73</td>
<td>Santo Domingo</td>
</tr>
<tr>
<td>El Salvador</td>
<td>20720</td>
<td>7066403</td>
<td>72</td>
<td>San Salvador</td>
</tr>
<tr>
<td>Grenada</td>
<td>344</td>
<td>90343</td>
<td>65</td>
<td>Saint George's</td>
</tr>
<tr>
<td>Guatemala</td>
<td>108430</td>
<td>13002206</td>
<td>69</td>
<td>Guatemala</td>
</tr>
<tr>
<td>Haiti</td>
<td>27560</td>
<td>8924553</td>
<td>57</td>
<td>Port-au-Prince</td>
</tr>
</tbody>
</table>

The Siboney were the first to inhabit the islands of Antigua and Barbuda in 2400 B.C., but Arawak Indians populated the islands when COLUMBUS landed on his second voyage in 1493. Early settlements by the Spanish and French were succeeded by the English who formed a colony in 1667. Slavery, established to run the sugar plantations on Antigua, was abolished in 1834. The islands became an independent state within the British Commonwealth of Nations in 1981.
Results Don’t Have to Be a List

Poland is an ancient nation that was conceived near the middle of the 10th century. Its golden age occurred in the 16th century. During the following century, the strengthening of the gentry and internal disorders weakened the nation. In a series of agreements between 1772 and 1795, Russia, Prussia, and Austria partitioned Poland amongst themselves. Poland regained its independence in 1918 only to be overrun by Germany and the Soviet Union in World War II. It became a Soviet satellite state following the war, but its government was comparatively tolerant and progressive. Labor turmoil in 1980 led to the formation of the independent trade union “Solidarity” that over time became a political force and by 1990 had swept parliamentary elections and the presidency. A “shock therapy” program during the early 1990s enabled the country to transform its economy into one of the most robust in Central Europe, but Poland still faces the lingering challenges of high unemployment, underdeveloped and dilapidated infrastructure, and a poor rural underclass. Solidarity suffered a major defeat in the 2001 parliamentary elections when it failed to elect a single deputy to the lower house of Parliament, and the new leaders of the Solidarity Trade Union subsequently pledged to reduce the Trade Union’s political role. Poland joined NATO in 1999 and the European Union in 2004. With its transformation to a democratic, market-oriented country largely completed, Poland is an increasingly active member of Euro-Atlantic organizations.
Conclusions

• CloudSearch represents the next stage in the democratization of search.
• You no longer need to be a search expert to be a search provider.
• As the number and variety of search applications increases, we should see an increase in the variety of search interfaces.
Questions?

To learn more about CloudSearch:
cloudsearch@amazon.com
https://aws.amazon.com/cloudsearch/

Thanks to Matt Amacker, Puneet Gupta, Asif Makhani, Brian Pinkerton, Joel Tesler