State-of-the-art search technology and future challenges

CTO Prof. Bjørn Olstad
Email: bjorn.olstad@fast.no

Fast Search and Transfer ASA
The Norwegian University of Science and Technology
**Company**
- Founded in ’97
- Sold Internet BU to Overture/Yahoo
- > 1000 customers
- #2 growing technology company in Europe 1998-2002

**Product**
- Enterprise Search Platform
- Extreme capabilities in
  - Scalability
  - Accuracy
  - Analytics
FAST’s Mission …

… Power the Most Challenging Information Retrieval Applications
FAST Research Strategy

- **Strategic innovation**
  - Securing long term viability through leading industrial strength engine for aggregation, mining and information discovery in structured/unstructured data repositories/feeds

- **Customer orientation**
  - Partnering with leading global companies to solve the biggest search challenges

- **University partnerships**
  - Strategic deep relations to:
    - Cornell: Fred Schneider, Trustworthy Computing
    - Penn. State: Lee Giles, Niche/Meta searching
    - Munich: Franz Guenthner, Linguistics
    - Trondheim/Tromsø: Algorithms/Architecture

- **EU 6th Framework research projects**
  - Currently 3 funded projects: Analytical search, Integration of search & case based reasoning, and grid based search architectures
Merging access to content and data

Unified Information Access

Content Access Tools: Search, Categorization, Analytics

Content Platform

Query, Reporting, and Analysis Tools

Database Platform

Source: IDC #30704, 2004: Changing the face of enterprise computing.
Solving the Information Crisis

Information Infrastructure — Content Management and Database Applications

Unstructured Media

Information

Structured Data
Search vs. Database Approach

SEARCH DOESN’T SUPPORT...

- Database transaction processing, rollback, ...
- Joins
- Extensive upfront schema modeling
- Pre-aggregation of values in data marts

... (therefore) SEARCH DO SUPPORT:

Scalability:
- 10-100 times more cost efficient data aggregation

Performance:
- 50-250 times lower search latencies

Text:
- Both unstructured & structured data

Intelligence:
- Ranking of results based on importance

Analytics:
- On-the-fly mining of meta data properties
Where We’re Going Now

• Applications that search can supercharge include:
  – Customer Relationship Management
  – Supply Chain Management
  – Business Intelligence
  – Market Intelligence
  – Research Support
  – Threat Detection
  – Anywhere data and unstructured text, speech or general volition collide
Web sites consist of two different kinds of pages

**Navigation pages**

- **ThinkPad Home Page**
  - **Purpose**: Move to next page
  - **User question**: Where do I go next?
  - **Traffic**: High
  - **Searches**: Broad queries

**Destination pages**

- **ThinkPad G40 Product Details**
  - **Purpose**: Provide information
  - **User question**: Is this what I wanted?
  - **Traffic**: Lower
  - **Searches**: Specific queries

Page types defined in “Information Architecture for the World Wide Web” by Louis Rosenfeld and Peter Morville, p. 139
Different queries require different approaches

**Broad queries**

Examples: notebook, laptop, thinkpad

Approaches: Keywords on page, Inbound links, Anchor text

**Specific queries**

Examples: thinkpad g40, 23887RU, 2388-7RU

Approaches: Keywords on page, Part numbers on page (including variations)
The Query – Document Relationship

General Queries

Problem Queries

Specific Queries

Content

Format

Reference

‘New York’

‘C source code quicksort’

‘HP printer driver LP 6j’
<table>
<thead>
<tr>
<th>MeSH term</th>
<th>chemical substance</th>
<th>publication</th>
<th>publication year</th>
<th>author</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>#</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>DNA</td>
<td>54</td>
<td>1998</td>
<td>K Pruitt</td>
</tr>
<tr>
<td>Animal</td>
<td>Genetic Markers</td>
<td>25</td>
<td>2000</td>
<td>J Weissenbach</td>
</tr>
<tr>
<td>Genome, Human</td>
<td>DNA Primers</td>
<td>8</td>
<td>1998</td>
<td>B Chowdhary</td>
</tr>
<tr>
<td>Genome</td>
<td>DNA, Mitochondrial</td>
<td>8</td>
<td>1999</td>
<td>L Fröncke</td>
</tr>
<tr>
<td>Chromeosome Mapping</td>
<td>DNA, Complementary</td>
<td>3</td>
<td>1996</td>
<td>H Scherthan</td>
</tr>
<tr>
<td>Support, Non-U.S. Govt</td>
<td>DNA Transposable Elements</td>
<td>11</td>
<td>1995</td>
<td>J McPherson</td>
</tr>
<tr>
<td>Base Sequence</td>
<td>Proteins</td>
<td>29</td>
<td>1999</td>
<td>I Gustavsson</td>
</tr>
<tr>
<td>Comparative Study</td>
<td>Plasmids</td>
<td>22</td>
<td>1998</td>
<td>G Elgar</td>
</tr>
<tr>
<td>Support, U.S. Govt, P.H.S.</td>
<td>DNA-Binding Proteins</td>
<td>13</td>
<td>2000</td>
<td>E Green</td>
</tr>
<tr>
<td>Human Genome Project</td>
<td>RNA, Viral</td>
<td>11</td>
<td>1997</td>
<td>T Raudsepp</td>
</tr>
<tr>
<td>Mice</td>
<td>DNA Probes</td>
<td>11</td>
<td>1998</td>
<td>M Kirby</td>
</tr>
<tr>
<td>Molecular Sequence Data</td>
<td>DNA, Viral</td>
<td>11</td>
<td>1998</td>
<td>C Fitzames</td>
</tr>
<tr>
<td>Sequence Analysis, DNA</td>
<td>DNA, Satellite</td>
<td>1</td>
<td>1995</td>
<td>A Billault</td>
</tr>
<tr>
<td>Databases, Factual</td>
<td>Muscle Proteins</td>
<td>1</td>
<td>1995</td>
<td>R Gyapay</td>
</tr>
<tr>
<td>grace</td>
<td>Membrane Proteins</td>
<td>1</td>
<td>1995</td>
<td>R Bud</td>
</tr>
<tr>
<td>Human Genome</td>
<td>Genome Res</td>
<td>54</td>
<td>1998</td>
<td>K Pruitt</td>
</tr>
<tr>
<td>Mamm Genome</td>
<td>Mamm Genome</td>
<td>25</td>
<td>2000</td>
<td>J Weissenbach</td>
</tr>
<tr>
<td>Law Hum Genome Rev</td>
<td>DNA, Mitochondrial</td>
<td>8</td>
<td>1999</td>
<td>B Chowdhary</td>
</tr>
<tr>
<td>Proc Natl Acad Sci U S A</td>
<td>DNA, Complementary</td>
<td>3</td>
<td>1996</td>
<td>L Fröncke</td>
</tr>
<tr>
<td>Genome</td>
<td>DNA Transposable Elements</td>
<td>11</td>
<td>1995</td>
<td>H Scherthan</td>
</tr>
<tr>
<td>J Mol Evol</td>
<td>Plasmids</td>
<td>22</td>
<td>2000</td>
<td>I Gustavsson</td>
</tr>
<tr>
<td>Virology</td>
<td>DNA-Binding Proteins</td>
<td>11</td>
<td>1997</td>
<td>G Elgar</td>
</tr>
<tr>
<td>Int J Parasitol</td>
<td>RNA, Viral</td>
<td>11</td>
<td>1997</td>
<td>E Green</td>
</tr>
<tr>
<td>FEMS Microbiol Rev</td>
<td>DNA Probes</td>
<td>1</td>
<td>1995</td>
<td>T Raudsepp</td>
</tr>
<tr>
<td>Rom J Physiol</td>
<td>DNA, Viral</td>
<td>1</td>
<td>1995</td>
<td>M Kirby</td>
</tr>
<tr>
<td>Hum Mol Genet</td>
<td>DNA, Satellite</td>
<td>1</td>
<td>1995</td>
<td>C Fitzames</td>
</tr>
<tr>
<td></td>
<td>Muscle Proteins</td>
<td>1</td>
<td>1995</td>
<td>A Billault</td>
</tr>
<tr>
<td></td>
<td>Membrane Proteins</td>
<td>1</td>
<td>1995</td>
<td>R Gyapay</td>
</tr>
<tr>
<td></td>
<td>DNA, Mitochondrial</td>
<td>11</td>
<td>1995</td>
<td>R Bud</td>
</tr>
<tr>
<td></td>
<td>DNA Transposable Elements</td>
<td>11</td>
<td>1995</td>
<td>R Bud</td>
</tr>
<tr>
<td></td>
<td>Proteins</td>
<td>29</td>
<td>1998</td>
<td>K Pruitt</td>
</tr>
<tr>
<td></td>
<td>Plasmids</td>
<td>22</td>
<td>2000</td>
<td>J Weissenbach</td>
</tr>
<tr>
<td></td>
<td>DNA-Binding Proteins</td>
<td>11</td>
<td>1997</td>
<td>H Scherthan</td>
</tr>
<tr>
<td></td>
<td>RNA, Viral</td>
<td>11</td>
<td>1997</td>
<td>J McPherson</td>
</tr>
<tr>
<td></td>
<td>DNA Probes</td>
<td>1</td>
<td>1995</td>
<td>I Gustavsson</td>
</tr>
<tr>
<td></td>
<td>DNA, Viral</td>
<td>1</td>
<td>1995</td>
<td>G Elgar</td>
</tr>
<tr>
<td></td>
<td>DNA, Satellite</td>
<td>1</td>
<td>1995</td>
<td>E Green</td>
</tr>
<tr>
<td></td>
<td>Muscle Proteins</td>
<td>1</td>
<td>1995</td>
<td>T Raudsepp</td>
</tr>
<tr>
<td></td>
<td>Membrane Proteins</td>
<td>1</td>
<td>1995</td>
<td>M Kirby</td>
</tr>
<tr>
<td></td>
<td>DNA, Mitochondrial</td>
<td>11</td>
<td>1995</td>
<td>C Fitzames</td>
</tr>
<tr>
<td></td>
<td>DNA Transposable Elements</td>
<td>11</td>
<td>1995</td>
<td>A Billault</td>
</tr>
<tr>
<td></td>
<td>Proteins</td>
<td>29</td>
<td>1998</td>
<td>K Pruitt</td>
</tr>
<tr>
<td></td>
<td>Plasmids</td>
<td>22</td>
<td>2000</td>
<td>J Weissenbach</td>
</tr>
<tr>
<td></td>
<td>DNA-Binding Proteins</td>
<td>11</td>
<td>1997</td>
<td>H Scherthan</td>
</tr>
<tr>
<td></td>
<td>RNA, Viral</td>
<td>11</td>
<td>1997</td>
<td>J McPherson</td>
</tr>
<tr>
<td></td>
<td>DNA Probes</td>
<td>1</td>
<td>1995</td>
<td>I Gustavsson</td>
</tr>
<tr>
<td></td>
<td>DNA, Viral</td>
<td>1</td>
<td>1995</td>
<td>G Elgar</td>
</tr>
<tr>
<td></td>
<td>DNA, Satellite</td>
<td>1</td>
<td>1995</td>
<td>E Green</td>
</tr>
<tr>
<td></td>
<td>Muscle Proteins</td>
<td>1</td>
<td>1995</td>
<td>T Raudsepp</td>
</tr>
<tr>
<td></td>
<td>Membrane Proteins</td>
<td>1</td>
<td>1995</td>
<td>M Kirby</td>
</tr>
<tr>
<td></td>
<td>DNA, Mitochondrial</td>
<td>11</td>
<td>1995</td>
<td>C Fitzames</td>
</tr>
<tr>
<td></td>
<td>DNA Transposable Elements</td>
<td>11</td>
<td>1995</td>
<td>A Billault</td>
</tr>
<tr>
<td></td>
<td>Proteins</td>
<td>29</td>
<td>1998</td>
<td>K Pruitt</td>
</tr>
<tr>
<td></td>
<td>Plasmids</td>
<td>22</td>
<td>2000</td>
<td>J Weissenbach</td>
</tr>
<tr>
<td></td>
<td>DNA-Binding Proteins</td>
<td>11</td>
<td>1997</td>
<td>H Scherthan</td>
</tr>
<tr>
<td></td>
<td>RNA, Viral</td>
<td>11</td>
<td>1997</td>
<td>J McPherson</td>
</tr>
<tr>
<td></td>
<td>DNA Probes</td>
<td>1</td>
<td>1995</td>
<td>I Gustavsson</td>
</tr>
<tr>
<td></td>
<td>DNA, Viral</td>
<td>1</td>
<td>1995</td>
<td>G Elgar</td>
</tr>
<tr>
<td></td>
<td>DNA, Satellite</td>
<td>1</td>
<td>1995</td>
<td>E Green</td>
</tr>
<tr>
<td></td>
<td>Muscle Proteins</td>
<td>1</td>
<td>1995</td>
<td>T Raudsepp</td>
</tr>
<tr>
<td></td>
<td>Membrane Proteins</td>
<td>1</td>
<td>1995</td>
<td>M Kirby</td>
</tr>
<tr>
<td></td>
<td>DNA, Mitochondrial</td>
<td>11</td>
<td>1995</td>
<td>C Fitzames</td>
</tr>
<tr>
<td></td>
<td>DNA Transposable Elements</td>
<td>11</td>
<td>1995</td>
<td>A Billault</td>
</tr>
</tbody>
</table>
Data Mining

Text Mining

Data Retrieval

Information Retrieval

Structured Data

Unstructured Data (Text)

Search (goal-oriented)

Discover (opportunistic)

Data Mining

Text Mining
BC-dynegy-enron-offer-update5
Dynegy May Offer at Least $8 Bln to Acquire Enron (Update5)
By George Stein
SOURCEc.2001 Bloomberg News BODY

```
Dynegy has to act fast," said Roger Hamilton, a money manager with John Hancock Advisers Inc., which sold its Enron shares in recent weeks. "If Enron can't get financing and its bonds go to junk, they lose counterparties and their marvelous business vanishes."

Moody's Investors Service lowered its rating on Enron's bonds to "Baa2" and Standard & Poor's cut the debt to "BBB." in the past two weeks.

......
```
Niels Bohr – Biography

Niels Henrik David Bohr was born in Copenhagen on October 7, 1885, as the son of Christian Bohr, Professor of Physiology at Copenhagen University, and his wife Ellen, née Adler. Niels, together with his younger brother Harald (the future Professor in Mathematics), grew up in an atmosphere most favourable to the development of his genius - his father was an eminentphysiologist and was largely responsible for awakening his interest in physics while still at school, his mother came from a family distinguished in the field of education.

After matriculation at the Gammelholm Grammar School in 1903, he entered Copenhagen University where he came under the guidance of Professor C. Christiansen, a profoundly original and highly endowed physicist, and took his Master’s degree in Physics in 1909 and his Doctor’s degree in 1911.
**Mark Shields: Stop-Dean Movement Stumbles**

Yesterday, Week, Month

**Kerry Airs Ads To Catch Up With Dean**

Yesterday, Week, Month

**Baghdad Hero Dismisses Ex-Nato Commander’s Presidential Bid**

Yesterday, Week, Month

**Poll: 50 Percent Of Voters Against Bush**

Yesterday, Week, Month

**Sen. John Kerry Fire’s Campaign Manager**

Yesterday, Week, Month

Related Companies
- Time Warner
- Associated Press
- Newsweek
- Digital Chicago
- Enron
- Chase Manhattan Mortgage
- Harry N Abrams
- State Farm Mutual Auto Ins
- Two Guys
### Meta-data

<table>
<thead>
<tr>
<th>Price</th>
<th>Color</th>
<th>Abstract</th>
<th>Names</th>
<th>Brands</th>
<th>Geo</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{10}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Auto-detected Entities

<table>
<thead>
<tr>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live Analytics™</td>
</tr>
<tr>
<td>Clustering</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

**Value**

- Viewed results

**Noise**

- Documents

**Live Analytics™**

- Value

**Analysis results**
Unstructured Data


We further identified that the RNA components are 335 ntRNA, 165 ntRNA, etc. and the RNA intermediate, 55 ntRNA, 1875 ntRNA and its decay intermediates, and demonstrated that RNA degradation is carried out in the degrader by RNase E cleavage of A-U rich single-stranded regions of mature 335 and 165 ntRNA (Proc. Natl. Acad. Sci. USA, 93:3157-61, 1996).

Our current investigations are aimed at the action of proteins in the degrader on RNA degradation and processing. First, we identified locations of immune-emic

Structured
PARIS (Reuters) - Venus Williams raced into the second round of the $11.25 million French Open Monday, brushing aside Blanka Lamsad, 6-3, 6-3, in 65 minutes.

The Wimbledon and U.S. Open champion, seeded second, breezed past the German on a blustery center court to become the first seed to advance at Roland Garros. "I love being here, I love the French Open and more than anything I'd love to do well here," the American said.

A first round loser last year, Williams is hoping to progress beyond the quarter-finals for the first time in her career.
The *InPerspective* ranking model

**Freshness**
- How fresh is the document compared to the time of the query?

**Completeness**
- How well does the query match superior contexts like the title or the url?
  - *Example: query=“Mexico”, Is "Mexico” or “University of New Mexico” best?*

**Authority**
- Is the document considered an authority for this query?
  - *Examples: Web link cardinality, article references, product revenue, page impressions, ...*

**Statistics**
- How well does the contents of this document overall match the query?
  - *Examples: Proximity, context weights, tf-idf, degree of linguistic normalization, ...*

**Quality**
- What is the quality of the document?
  - *Examples: Homepage?, Entry point to product group?, Press release?, ...*
Query: Do you have a LCD monitor under $900?

Tokenizer → Spellcheck → Phrasing → Anti-phrase → Normalize

Under $900? → LCD monitors TFT monitor Flat TV Plasma TV

YES! X = LCD monitor

NLQ → Lemmas Synonyms Thesaurus PLUG-IN BUY( X )

Use “Product” collection
Rank profile = “Profit margin”

Modified query
Virgilio’s Results

Results so far...

After New Search Launch:
- +27% Avg. traffic growth
- +12% Avg. users growth
- +12% Relevance Index vs Google
- Market leader in € value
- Sole competitor vs. usage leader
• Search engines can do more than just search…
  – Unified information access solution for digital libraries
  – Open, scalable and modular architecture: Allows for customization
  – Adapts to content and queries
  – Powerful data discovery, navigation, and visualization

• Many exciting technology developments to come
  – More advanced content and query analysis
  – Adaptive, personalized query- & content-sensitive matching
  – Dynamic result set presentation, navigation, discovery, visualization
  – Federation across external content applications