Multilingual Information Access and Available Resources

Gregory Grefenstette, Exalead

MultiLingual Information Access Technology Transfer Day
World Business Alliance Center, Berlin, Germany
8 December 2009
Search Based Information Access Technology

- Founded in 2000, two former AltaVista engineers
- Headquarters in Paris, France
- 180 employees worldwide
- 30% in R&D
- 220 customers worldwide
- Revenue 2008: +80%
- Major industries: B2B, OEM, B2C
- Major partners

120 million unique visitors/month already access information through Exalead technology

Exalead’s proven scalability, with one user today already processing almost a petabyte of data and updating 2% of this amount daily in real-time.
Exalead appears for the fourth consecutive year in **Gartner’s Magic Quadrant and is a Challenger company in the 2009** Gartner Information Access Magic Quadrant.

IDC Content Technologies Group: “**Exalead is disruptive because the company has moved aggressively from Web search to enterprise search, and now to information access. The firm’s technology makes it possible to integrate structured and unstructured content in a unique way to address mission-critical applications in areas such as extended business intelligence, customer support, compliance, and many others.**”
On Saturday, in Newcastle, England, two fighters will clash for the WBA light-welterweight world title. The December 5 bout, on the surface, seems much like any other fight for a globally-recognised belt, but there is an intriguing religious sub-plot to the clash. This is because reigning world champion Amir Khan is a practicing Muslim while his mandatory opponent Dmitriy Salita is an Orthodox Jew, a fact which has meant certain factors needed to be considered. For a start, the fight is scheduled for after sundown to accommodate observance of the Jewish Sabbath by Salita, a 27-year-old Ukrainian-turned American, and a large group of Orthodox Jews from London who had already planned to attend the fight to support him. A documentary film-maker is even flying in a group of Jewish and Muslim teenagers for a cross-cultural experience, in a bid to underline how this occasion can act as an event to
A few Search Engine Fields capture most semantics

- Field types
  - Text
  - Dates
  - Numbers
  - Categories
8 billion web pages
2 billion images
17 million videos
Hubble -- Out of the ordinary...out of this world.
Search all of HubbleSite HOME NEWSCENTER GALLERY HUBBLE DISCOVERIES HUBBLE [...] NEWSCENTER hubble news All the Hubble news and pictures, from the latest discoveries to the earliest observations.
hubblesite.org - Cached - Bookmark

Main Hubble Page
For instructions, click here + NASA HOME + Contact Hubble Program Named after the trailblazing astronomer Edwin P. Hubble (1889-1953), the Hubble Space Telescope (HST) is a large space...
hubble.nasa.gov - Cached - Bookmark

The Hubble Heritage Project Website
Last WFPC2 Image New: Join us on Facebook and Flick!
heritage.stsci.edu - Cached - Bookmark

Esa Science & Technology: Hubble
During its lifetime Hubble has become one of the most important science projects ever. NEWS Hubble captures rare Jupiter collision [heic0909] Commissioning Hubble - preparing for science ...
hubble.esa.int - Cached - Bookmark

The European Hompage For The Nasa/esa Hubble Space Telescope
hubble.org - Cached - Bookmark
Flattened facets
### Tabular facets

#### En Cours

<table>
<thead>
<tr>
<th>Agence départ</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEFCO LE HAVRE</td>
<td>259 97.07 %</td>
</tr>
<tr>
<td><em><strong>À SUPPRIMER</strong></em></td>
<td>9 2.92 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type de transport</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Livraison maritime</td>
<td>264 52.9 %</td>
</tr>
<tr>
<td>Usine</td>
<td>191 38.27 %</td>
</tr>
<tr>
<td>Transit sur Centre</td>
<td>44 8.81 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type d'opération</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Livraison maritime</td>
<td>264 52.9 %</td>
</tr>
<tr>
<td>Usine</td>
<td>191 38.27 %</td>
</tr>
<tr>
<td>Transit sur Centre</td>
<td>44 8.81 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agence arrivée</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHBA</td>
<td>264 85.71 %</td>
</tr>
<tr>
<td>GEFCO LE HAVRE</td>
<td>35 11.36 %</td>
</tr>
<tr>
<td><em><strong>À SUPPRIMER</strong></em></td>
<td>9 2.92 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type de véhicule</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>VN</td>
<td>499 100.0 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Destination</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPN</td>
<td>499 100.0 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usine de production</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>(102) Sochaux</td>
<td>190 38.46 %</td>
</tr>
<tr>
<td>(130) Poissy</td>
<td>139 28.13 %</td>
</tr>
<tr>
<td>(120) Rennes</td>
<td>101 20.44 %</td>
</tr>
<tr>
<td>(368) Vigo</td>
<td>42 8.5 %</td>
</tr>
<tr>
<td>(108) Mulhouse</td>
<td>11 2.22 %</td>
</tr>
<tr>
<td>377</td>
<td>9 1.82 %</td>
</tr>
<tr>
<td>(367) Villaverde</td>
<td>2 0.4 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marque</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEUGEOT</td>
<td>351 70.34 %</td>
</tr>
<tr>
<td>CITROEN</td>
<td>143 28.65 %</td>
</tr>
<tr>
<td>TOUTES MARQUES</td>
<td>5 1.0 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ligne de produit</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>308</td>
<td>191 36.27 %</td>
</tr>
<tr>
<td>207</td>
<td>141 28.25 %</td>
</tr>
<tr>
<td>C5 2</td>
<td>85 17.03 %</td>
</tr>
<tr>
<td>C4 PICASSO</td>
<td>42 8.41 %</td>
</tr>
<tr>
<td>407</td>
<td>10 2.0 %</td>
</tr>
<tr>
<td>C4</td>
<td>10 2.0 %</td>
</tr>
<tr>
<td>206</td>
<td>9 1.8 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statut blocage</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non bloqué</td>
<td>498 99.79 %</td>
</tr>
<tr>
<td>Bloqué</td>
<td>1 0.2 %</td>
</tr>
<tr>
<td>___ Exp</td>
<td>1 100 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Centre de gestion</th>
<th>Voir le graphique</th>
</tr>
</thead>
<tbody>
<tr>
<td>LE HAVRE</td>
<td>299 97.07 %</td>
</tr>
<tr>
<td>U.K. PC TRANSITAIRE</td>
<td>9 2.92 %</td>
</tr>
</tbody>
</table>
Semantics comes from Databases
Text-based Semantic Modules

autosuggestion
language identification
related terms
related queries
spell checker
faceted search

lemmatisation
synonyms
phonetic
cross language
Ontology Matcher

phonetic
lemmatisation
stopwords
sentiment analysis
named entity

QueryMatcher
FastRules
HTMLRelevantContextExtractor
Categorizer
Clusterer

Data Sources

INDEX
Auto completion of user query as they type
Synonym, Example

DB administrator

is defined as synonym of

Database Administrator

This synonymy can be in one direction or both ways
54 languages are currently detected.
Cloning

Cloning: Hopefully, this will turn into a full fledged tool to recompute the physical map and assess its reliability. At present, it does very little. This option is not linked in release 1.4. Select project. This will look for a binary file $ACEDB_DATA/subseton/whatever_you_enter.bnd. If it does not exist, the program will try

Source: AceDB; public websites

TAR CLOTHING

Two stages (see figure 1): In the first stage a targeting vector is created, by cloning a 500bp PCR product into the Xho-I site of pEB2. The PCR product Telomere Capture URA3 pEB2 6548 bp Xho I TEL Cla I 500bp PCR product is cloned into Xho-I Site. Correct orientation has: CEN TEL. Plasmid is linearised by digestion.

Source: Library; public websites

S.scrofa

and also the name of the BAC from which each BES was produced. A BAC clone name e.g. CH243-196P11, provides a link to further information containing, where possible, as shown in the 'Pig BAC ends' track in human ensembl. d, porcine sequenced clone CH242-196P11 (CR9... contains the porcine homologue to GNMT. e, pig clone CH242-196P11 (CR9... which
Cross Language Information Retrieval

afghanistan

Show tags
Show Time Line

October 12 2009...参议院军事委员会主席基文说在阿富汗制定更好的战略以增加...

August 25 2009...一些专家阿富汗总统大选后的形势受到各方高度关切...

December 1 2009...爱尔兰总理美国和阿富汗总统卡尔扎伊进行了...

December 1 2009...计划于日本时间明天宣布新的阿富汗战略在此之前他一直是...

November 30 2009...如果是有好处的话他们有什么样的话阿富汗恐怖活动他们只...
Spell Checker
Cluster analysis, or automatic classification, is a multivariate statistical technique... henceforth a GA, for document clustering.

**FastRules**

- **100 000 queries with boolean large multi word expression matching 1 Mb / sec / server**

- **FastRulesAdvantages**
  - Very fast rules categorization
  - Support huge set of rules (> 100k rules)
  - Regular expression on chunk (for example a regular expression on an url, title or text)
  - Available as a standalone component (Exa and Java)

- **Drawbacks**
  - Limited syntax
    (only AND/OR/NOT operators are supported)
  - Rules must be compiled before usage

**Jones, G. et al. Non-hierarchic document clustering** Willett

Introduction

Cluster analysis, or automatic classification, is a multivariate statistical technique... henceforth a GA, for document clustering.

04 nov. 2003 - 20k

**MyCategory**: MachineLearning


**Personnalité**:

Gareth Jones, Peter Willett, Edward Arnold, Van Nostrand Reinhold

**Lieu**:

Berlin, Duran, Sheffield, New York, London

**Organisation**:

British Library, Wellcome, Everitt, American Society, ACM, Springer
Ontology Matcher
add an alternative indexing form
to a string span

For example, we can use the ontology matcher so that searching any postal code will also search for any cities matching that code.

2 millions multi word expression matching: 20 Mb / sec / server
Michelle Obama
See also: Michelle Robinson's family tree [...] 2008 Presidential election Campaigning for Barack Obama [...] conservative columnist Michelle Malkin about ...

Oprah Winfrey’s endorsement of Barack Obama
Endorsing Barack Obama [...] vote for Obama, if only because they thought her endorsement made Obama more [...] It made Obama more than a 'Black' candidate.

Marian Shields Robinson
[... ] Lady of the United States Michelle Obama. She is the mother-in-law of President Barack Obama and the only surviving grandparent of Malia and Sasha Obama.
SentimentAnalyzer

Blue Smoke  🌟🌟🌟🌟
116 E 27th St, New York, NY 10016, USA
Phone: (212) 447-7733
Subway: 6 at 28th St
Website: http://www.bluesmoke.com - Book Online!
Kind of food: Barbecue, American, American Traditional, Southern & Soul, American (Traditional)

Overview

Review

"Forget what the 'cue quibblers say. Blue Smoke is an ingenious Manhattan barbecue joint with a woody barroom full of red-vinyl booths, a skylit dining room, and a hot jazz club downstairs. True, some of the barbecue isn't there yet, but each visit shows improvement, as pit master Kenny Callaghan breaks in those Missouri-made smokers. Salt-and-pepper beef ribs are terrific, so are sides and desserts, and as far as pit-smoked pork and salmon go, what's the problem? This ain't east Texas barbecue, Texas, after all."

Most viewed comment

"Went to Blue Smoke tonight. The waitstaff was friendly, even if the last sip of my beer was whisked away without so much as an "Are you finished with that?" Had the barbeque ribs, with a side of fries, mac & cheese and hush-puppies. The Individuals that come with the hush-puppies makes this a real specialty. The mac & cheese was creamy, to put it mildly. Very tasty, but for the hint of chemically-vibes like after-taste. The fries - excellent; they arrived fresh, hot, with a crispy exterior (no better than these little twisters) and with the genuine taste of potato without... More >>"

From New York Magazine

People sentiments

reasonable, okay, spicy, inspired, hot, baked, overpriced, juicy, hard, grilled, bizarre, famous, impressed, bold, impressive, dry, amazing, consistent, negative, expensive, superb, efficient, endless, thin, meaty, juicy, saucy, available, genuine, cute, typical, picky, yummy, good, popular, outer, tastiness, pretty, incredible, obvious, like, unreasonable, low, comfortable, pricey, savvy, boring, Ok, stellar, affordable, indifferent, diagnosed, quality, sorry, unbelievable, vinegary, returning, small, delicious, unbelievable, excited, refreshing, usable, nasty, traditional, polite, serviceable, exceptional, tasty, incredible, huge, fond

From New York Magazine

Restminer

Blue Smoke

Blue Smoke

116 E 27th St, New York, NY 10016
Website: http://www.bluesmoke.com - Book Online!
Kind of food: Barbecue, American, American Traditional, Southern & Soul, American (Traditional)

Overview

Review

"Forget what the 'cue quibblers say. Blue Smoke is an ingenious Manhattan barbecue joint with a woody barroom full of red-vinyl booths, a skylit dining room, and a hot jazz club downstairs. True, some of the barbecue isn't there yet, but each visit shows improvement, as pit master Kenny Callaghan breaks in those Missouri-made smokers. Salt-and-pepper beef ribs are terrific, so are sides and desserts, and as far as pit-smoked pork and salmon go, what's the problem? This ain't east Texas barbecue, Texas, after all."

Most viewed comment

"Went to Blue Smoke tonight. The waitstaff was friendly, even if the last sip of my beer was whisked away without so much as an "Are you finished with that?" Had the barbeque ribs, with a side of fries, mac & cheese and hush-puppies. The Individuals that come with the hush-puppies makes this a real specialty. The mac & cheese was creamy, to put it mildly. Very tasty, but for the hint of chemically-vibes like after-taste. The fries - excellent; they arrived fresh, hot, with a crispy exterior (no better than these little twisters) and with the genuine taste of potato without... More >>"

From New York Magazine

People sentiments

reasonable, okay, spicy, inspired, hot, baked, overpriced, juicy, hard, grilled, bizarre, famous, impressed, bold, impressive, dry, amazing, consistent, negative, expensive, superb, efficient, endless, thin, meaty, juicy, saucy, available, genuine, cute, typical, picky, yummy, good, popular, outer, tastiness, pretty, incredible, obvious, like, unreasonable, low, comfortable, pricey, savvy, boring, Ok, stellar, affordable, indifferent, diagnosed, quality, sorry, unbelievable, vinegary, returning, small, delicious, unbelievable, excited, refreshing, usable, nasty, traditional, polite, serviceable, exceptional, tasty, incredible, huge, fond

From New York Magazine
No document matched your search
"the with in of "

Suggestions:
- Check the spelling of your terms
- Try different or more general terms
- Try to remove some terms
- Check the syntax of your request
- Frequent terms have to be quoted to be searched for
- You expected to find a certain site? Try to add it to the Exalead index!
**Predict** category of a new document using an existing training dataset (for example: dmoz)
Creamer shoots 60, leads Farr by 5 strokes (AP)

AP - Paula Creamer knew she was playing well. She was surprised just how well.

On This Topic: Creamer shoots 60, leads Farr by 5 strokes (AP) (Yahoo! News); Creamer shoots 60 in Jamie Farr Classic (PA SportsTicker) (Yahoo! News); Creamer shoots 60, leads Farr by 5 strokes (AP) (Yahoo! News); Creamer shoots 60 in 1st round of Farr (AP) (Yahoo! News); Creamer shoots 60 in 1st round of Farr (AP) (Yahoo! News); More on this topic

Ronaldo backs Blatter’s view of ‘modern slavery’

Cristiano Ronaldo has backed FIFA president Sepp Blatter’s assessment of ‘modern slavery’ with Manchester United.

On This Topic: Ronaldo backs Blatter’s view of ‘modern slavery’ (Fox Sports); Cristiano Ronaldo hints at Manchester United exit (AP) (Yahoo! News); Ronaldo backs Blatter ‘slave’ outburst (AP) (Yahoo! News); Ronaldo backs Blatter ‘slave’ outburst (AP) (Yahoo! News); Sepp Blatter in spotlight after ‘slave’ gaffe (Telegraph); More on this topic

Former PGA Tour player Souchak dies at 81 (Reuters)

American Mike Souchak, a 15-times winner on the PGA Tour who appeared on two U.S. Ryder Cup teams, died on Thursday after a heart attack, the Tour said.

On This Topic: Former PGA Tour player Souchak dies at 81 (Reuters) (Yahoo! News); Golf-Former PGA Tour player Souchak dies at 81 (Reuters) (Yahoo! News); Former PGA Tour champ Souchak dies (PA SportsTicker) (Yahoo! News); Souchak, former PGA Tour golfer, dies at 81 (AP) (Yahoo! News); Former tour record holder Souchak dies at 81 (Sportsline); More on this topic
Processing Pipelines

**Get the most out of your content**

- **Document Conversion**
  - Over 300+ file format

- **Document Processor (Others & Custom)**
  - Date normalization, Metadata processing

- **Language Detection**
  - 54 languages supported

- **Tokenizer**
  - Split words, detect end of phrase, ...

- **Part of Speech**
  - Detect word types (noun, verb, ...)

- **Lemmatization / Stemming**
  - Determine lemma and stem of the word

**Synonym**
- Associate Synonyms for predefined tokens

**Built-in Named Entity**
- Extract People, Location, Organization, ...

**Custom Named Entity (Ontology, Thesaurus)**
- Based on your own thesaurus

**Related Terms**
- Determine important Keywords

**Categorization**
- Use to determine similar documents, documents of same subject matters, ...

**Semantic Processors (Others & Customs)**
- Others Semantic processors such as **Sentiment Analysis**, ...

**Indexing**

Indexing Side
Search Pipelines

Understand the user query & enhance results

- **Language Detection**: 54 languages supported (used to choose the right tokenizer)
- **Parsing (Tokenizer)**: Split words, detect end of phrase, ...
- **Synonym Expansion**: Expanded query with user-defined synonym
- **Lemmatization / Phonetic**: Determine lemma and stem of the word. Apply Phonetization algorithms
- **Query Rewriting (Regexp, ...)**: Rewrite special expressions such as word*
- **Transformation into index query**: Query is rewritten to be comprehensible by the index

**Index**

- **Return results**: Index return results set matching user query and security rights
- **Ranking**: Rank results using density, text scoring and ranking formula
- **Related Terms**: Extract Related Terms from the result set
- **Summary**: Determine Summary to be displayed for each hit
- **Highlighting**: Highlight the words that are matching the user query

Search Side
A Defining Moment Nears for President

Six months into his administration, President Obama is at a pivotal moment. He has pushed through a $787 billion economic stimulus package, bailed out Wall Street and, on Tuesday, managed to beat the defense industry in the Senate, which voted to kill a high-profile fighter jet program. On Wednesday night Mr. Obama addresses the nation in a prime-time news conference as the public, and lawmakers, are growing skittish over his next big plan, to remake the American health care system.
Six months into his administration, President Obama is at a pivotal moment. He has pushed through a $787 billion economic stimulus package, bailed out Wall Street and, on Tuesday, managed to beat the defense industry in the Senate, which voted to kill a high-profile fighter jet program. On Wednesday night Mr. Obama addresses the nation in a prime-time news conference as the public, and lawmakers, are growing skittish over his next big plan, to remake the American health care system.
Six months into his administration, President Obama is at a pivotal moment.
### Sources of semantic information

<table>
<thead>
<tr>
<th>Web 2.0 Other Users</th>
<th>Ontologies</th>
<th>Natural Language Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Similar Queries</td>
<td>• Autosuggestion</td>
<td>• Language ID</td>
</tr>
<tr>
<td>• Related Terms</td>
<td>• Categorization</td>
<td>• Site type</td>
</tr>
<tr>
<td>• Spelling correction</td>
<td>• Facets</td>
<td>• Phonetics</td>
</tr>
<tr>
<td></td>
<td>• Named Entities</td>
<td>• Sentiment Analysis</td>
</tr>
<tr>
<td></td>
<td>• Ontology matching</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Synonyms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Cross Lingual Search</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Database Offloading</td>
<td></td>
</tr>
</tbody>
</table>
Real World Sources

• Typed sources (ontologies)
  – Wikipedia
    • Categories
    • Infoboxes
    • Multilingual sidebar
  – LinkedData
    • Flattened RDF triples

• Linguistic Resources
  – WordGumbo
  – ELDA
  – Europoean Projects
    • Clarins
    • FlareNet
AutoSuggestion Example

Auto completion of user query as they type

Possible MI source: Wikipedia entries

- **The Mozart Project**
- **Wolfgang Amadeus Mozart**
  - [http://w3.rz-berlin.mpg.de/cmp/mozart.html](http://w3.rz-berlin.mpg.de/cmp/mozart.html)
- **The Mozart Programming System**
DB administrator is defined as synonym of Database Administrator. This synonymy can be in one direction or both ways.

Possible ML source: EuroWordnets #REF Wikipedia.
Phonetic Example

Possible ML source: CMU phonetic dictionary
54 languages are currently detected

Possible ML source:
- EuroParl
- Wikipedia
- UDHR
- Google News
Possible ML source:

WordGumbo
CLARIN
ELDA
FlareNet
Cross Language Information Retrieval

Possible ML source:

Wikipedia

WordGumbo
Spell Checker

Possible ML source:
Ispell, ASpell
Cluster analysis, or automatic classification, is a multivariate statistical technique... henceforth a GA, for document clustering.

04 nov. 2003 - 20k

ManualCategorization: MachineLearning
Personnalité:
Gareth Jones, Peter Willett, Edward Arnold, Van Nostrand Reinhold
Lieu:
Berlin, Duran, Sheffield, New York, London
Organisation:
British Library, Wellcome, Everitt, American Society, ACM, Springer


Complexity depends of the number of rules (Conceptually, you should imagine we test each of the rule on the document)
Performance are not good if you have a lot of rules (> 100)
Regular expression on words (like the query parser)
Not available as a standalone component. Only available in indexing pipeline because it has a lot of interaction with query-parser, gdict, ...

```
<QueryMatcher annotationName="MyCategory" >
  <QueryList query="clustering AND (algorithm OR analysis OR learning)" annotationValue="MachineLearning">  
  <QueryList query="clustering AND "quote;load balancing"quote;" annotationValue="Hardware/Cluster"> 
  </QueryList>
  ...
</QueryMatcher>

<AnnotationMapping name="MyCategory" >
  <CategoryContentTarget retrievable=True indexField="corporate" categoryRoot="Top/ManualCategorization" categoryAppend=True form="exact" />
</AnnotationMapping>
```
Cluster analysis, or automatic classification, is a multivariate statistical technique ... henceforth a GA, for document clustering.

04 nov. 2003 - 20k

**MyCategory:** MachineLearning

**Source:** Vie société : Documentation : Press : Exalead EN-FR Press excerpts : Old - News Articles : Articles de recherche : NeuralNetworks :clustering

**Personnalité:**
Gareth Jones, Peter Willett, Edward Arnold, Van Nostrand Reinhold

**Lieu:**
Berlin, Duran, Sheffield, New York, London

**Organisation:**
British Library, Wellcome, Everitt, American Society, ACM, Springer

---

FastRules

```xml
<FastRulesDefinition xmlns="exa:com.exalead.mot.components.fastrules" catName="MyCategory">
    <Category value="MachineLearning">
        <Rule value="text: &quot;clustering&quot; AND text: &quot;algorithm&quot; OR text: &quot;analysis&quot; OR text: &quot;learning&quot;" exceptionRule="false"/>
    </Category>
    <Category value="Hardware/Cluster">
        <Rule value="text: &quot;clustering&quot; AND text: &quot;load balancing&quot;" exceptionRule="false"/>
    </Category>
</FastRulesDefinition>
```

---

**Possible ML source:**

**Wikipedia**

(only AND/OR/NOT operators are supported)

Rules must be compiled before usage

---

100 000 queries with boolean large multi word expression matching 1 Mb/sec/server

---

FastRulesDefinition

add a category to a document matching a compiled regular expression
Ontology Matcher
add an alternative indexing form
to a string span

For example, we can use the ontology matcher so that searching any postal code will also search for any cities matching that code.

2 millions multi word expression matching:
20 Mb / sec / server
RulesMatcher
Named Entities

Possible ML source:

Wikipedia InfoBoxes
LinkedData
**SentimentAnalyzer**

**Overview**

**Review**

“Forget what the ‘cue quibblers say97blue smoke is an ingenious manhattan barbecue joint with a woody barroom full of red-vinyl booths, a skylit dining room, and a hot jazz club downstairs; true, some of the barbecue isn’t there yet, but each visit shows improvement, as pit master kenny callaghan breaks in those missouri-made smokers. salt-and-pepper beef ribs are terrific, so are sides and desserts, and as far as pit-smoked foie gras and salmon goes, what’s the problem? this ain’t east brentwood, texas. a lot.”

**Most viewed comment**

“Went to blue smoke tonight. The waitstaff was friendly, even if the last sip of my beer was whisked away without so much as an ‘are you finished with that?’ Had the baby back ribs, with a side of fries, mac & cheese and hush-puppies. The remoulade that comes with the hush-puppies makes this dish a real specialty. The mac & cheese was creamy, to put it mildly. Very tasty, but for the hint of chemically-volatilizes-as-affect-false. The fries - excellent. They arrived fresh, hot, with a crispy exterior (no batter on these little critters) and with the genuine taste of potato. (More >>>)

**Blogged entries about Blue Smoke**

- Restaurant Week '08: Chase Away the Blues At Blue Smoke (New York City's Hottest Restaurant Deals)
- Summer Restaurant Week 2007 Pick: Blue Smoke (New York City's Hottest Restaurant Deals)
- Hill County to Challenge Blue Smoke, RUB on Their Own Turf (Grub Street)
- Restaurant Week Picks: Winter 2006 (Part 2) (New York City's Hottest Restaurant Deals)
- The Big Apple BBQ Block Party Demands a Lot From Its Vendors (Grub Street)
- Meyer Makes a Move on Grill Field (Grub Street)
- Robbie Richter and Hill Country Past Wars (Grub Street)
- B. R. Guest's New BBQ: Will Wildwood Succeed? (Grub Street)
- Heroic Blogger Saves Burger Lovers From Winter of Discontent (Grub Street)
- Profile: Midtown Lunch's "Yankees" (a Big Apple BBQ Fast Pass Giveaway) (Midtown Lunch)
- PROFILE: Midtown Lunch's "Chris" (Midtown Lunch)
- Daley May's BBQ Cart... the best-scratch that-ONLY BBQ in Midtown (Midtown Lunch)

**Details**

- Payment options: AMEX, Diners Club, Discover, MasterCard, Visa

Possible ML source: EuroWordnets
Possible ML source: Wikipedia InfoBoxes
Predict category of a new document using an existing training dataset (for example: dmoz)
Clusterer

Possible ML source:

SnowBall Stemmers
Why don’t Applications use research results?

• Licensing
• API
• Maintenance
### Missing : 260 word lists

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Surface</th>
<th>Lemma</th>
<th>POS</th>
<th>translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>acquiescassions</td>
<td>acquiescer</td>
<td>V</td>
<td>acquiesce, consent, approve</td>
</tr>
<tr>
<td>540</td>
<td>acquiesçons</td>
<td>acquiescer</td>
<td>V</td>
<td>acquiesce, consent, approve</td>
</tr>
<tr>
<td>77</td>
<td>acquiesçâmes</td>
<td>acquiescer</td>
<td>V</td>
<td>acquiesce, consent, approve</td>
</tr>
<tr>
<td>255</td>
<td>acquiesçât</td>
<td>acquiescer</td>
<td>V</td>
<td>acquiesce, consent, approve</td>
</tr>
<tr>
<td>6</td>
<td>acquiesçâtes</td>
<td>acquiescer</td>
<td>V</td>
<td>acquiesce, consent, approve</td>
</tr>
<tr>
<td>426000</td>
<td>acquirent</td>
<td>acquérir</td>
<td>V</td>
<td>acquire, purchase, buy</td>
</tr>
<tr>
<td>5290000</td>
<td>acquis</td>
<td>acquérir</td>
<td>V</td>
<td>acquire, purchase, buy</td>
</tr>
</tbody>
</table>

- And then multiword expressions
Conclusions

• For information access
  – typing needed (facets=types=categories)
    • connecting structured and unstructured

• Language related semantics intervenes
  – Search time, index access, index creation

• Resources needed
  – Language specific ontologies
  – Cross language terminology translation
  – Language normalisation
  – Language id, phonetics, tokenisation, ....