## **Open Source Search**

Andreas Pesenhofer

max.recall information systems GmbH Künstlergasse 11/1 • A-1150 Wien • Austria

ICIC, October 2013



# max.recall information systems

- max.recall is a software and consulting company enabling enterprises to capitalize on the hidden value in the rapidly growing amount of textual data
- Customized Solutions for
  - Intelligent data analytics
  - Vertical search
- Products and Services
  - quantalyze: quantity analytics technology
  - smart.coder: open-ended question coding tool for market researchers
- Founded 2010 and located in Vienna, Austria
- Operates worldwide with int'l customers from sectors such as IP, market research, news and media, IT services



## **Recall and precision**

- Recall •
  - Percent of relevant documents (items) returned
  - 50 good answers in system, 25 returned = 50% recall —
- Precision ٠
  - Percent of documents returned that are relevant
  - 100 returned, 25 are relevant = 25% precision —
- Ideal is 100% recall and 100% precision: return all relevant documents and ۲ only those
- 100% recall is easy return all documents, but precision is low, relevant ٠ documents can't be found
- Need adequate recall & enough precision for the task that will vary by ۲ application (data & users)



## How to get good recall

- Collect, index and search all the data
  - Check for missing or corrupt data
- Index everything
  - Search everything ... limit results by category AFTER the search (clustering/faceting)
- Normalize the data
  - Convert to lower case, strip/handle special characters, stemming, ...
- Use spell-checking, synonyms to match users' vocabulary with content
  - Adaptive spell-checking, application-specific synonyms
- Light (or real) natural language processing for abstract concepts

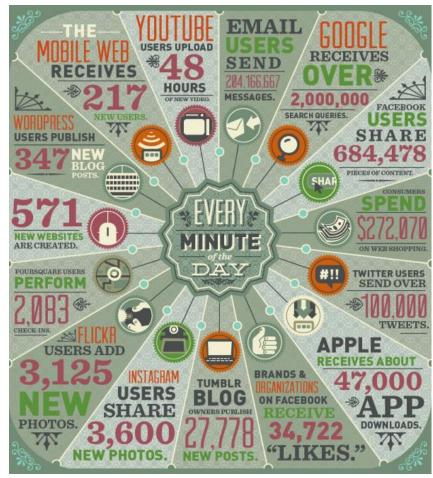


## How to get good precision

- Term frequency (TF) more occurrences of query terms is better
- Inverse document frequency (IDF) rarer query terms are more important
- Phrase boost query terms near each other is better
- Field boost where the query term is in doc matters (e.g., in 'title' better)
- Length normalization avoid penalizing short docs
- Recency all things being equal, recent is better
- Authority items linked to, clicked on or bought by others may be better
- Implicit and explicit relevance feedback, more-like-this expand query
- Clustering/faceting intent is not specific
- Lots of data



## Every minute ....

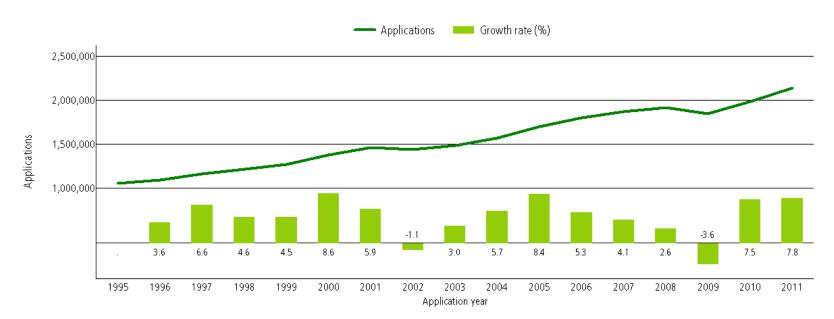


http://www.domo.com/



## **Groth of patent applications**

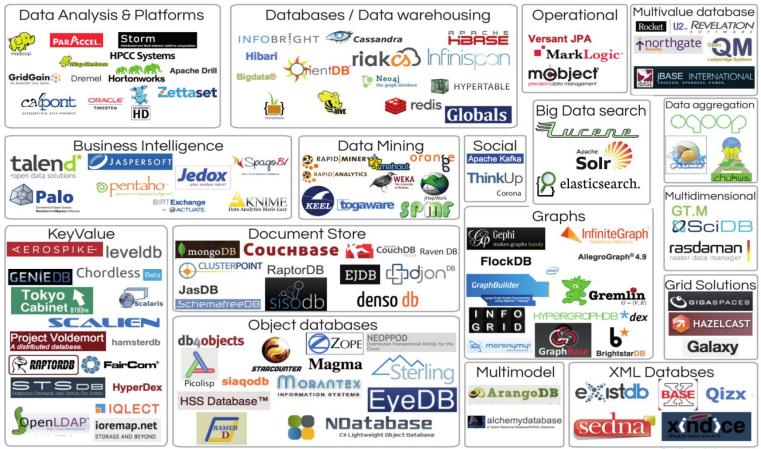
Trend in patent applications worldwide



Note: World totals are WIPO estimates covering around 125 patent offices (see Data Description)- These estimates include direct applications and PCT national phase entry data-Source: WIPO Statistics Database, October 2012



## **Big Data Open Source Tools**



Created by: www.bigdata-startups.com



## **Apache Lucene**



Apache Lucene<sup>™</sup> is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

- Scalable, High-Performance Indexing
  - over 150GB/hour on modern hardware
  - small RAM requirements only 1MB heap
  - incremental indexing as fast as batch indexing
  - index size roughly 20-30% the size of text indexed



## Apache Lucene (2)



- Powerful, Accurate and Efficient Search Algorithms
  - ranked searching best results returned first
  - many powerful query types: phrase queries, wildcard queries, proximity queries, range queries and more
  - fielded searching (e.g. title, author, contents)
  - sorting by any field
  - multiple-index searching with merged results
  - allows simultaneous update and searching
  - flexible faceting, highlighting, joins and result grouping
  - fast, memory-efficient and typo-tolerant suggesters
  - pluggable ranking models, including the Vector Space Model and Okapi BM25
  - configurable storage engine (codecs)



## Apache Lucene (3)



- Cross-Platform Solution
  - Available as Open Source software under the Apache License Lucene in both commercial and Open Source programs
  - 100%-pure Java
  - Implementations in other programming languages available, the index is compatible
- Apache Lucene 4.5.0 was released on October 5<sup>th</sup>, 2013.



## Apache SOLR



- Apache SOLR is an open source enterprise search platform from the Apache Lucene project.
- major features:
  - full-text search
  - hit highlighting
  - faceted search
  - dynamic clustering
  - database integration
  - handling of rich documents (e.g., Word, PDF)
  - providing distributed search and index replication, Solr is highly scalable.
- Apache SOLR 4.5.0 was released on October 5<sup>th</sup>, 2013.



## elasticsearch



- elasticsearch is a distributed, RESTful, open source search server based on Apache Lucene. It is developed by Shay Banon and is released under the terms of the Apache License.
- major features:
  - fully supports the near real-time search of Apache Lucene
  - cluster setup needs no additional software
  - features of Lucene are made available through the JSON and Java API
  - JSON in / JSON out (and YAML)
- elasticsearch 0.90.5 was released on September 17<sup>th</sup>, 2013, based on Lucene 4.4.



## **All Time Top Committers**

Name	12 Month Commits	All Time Commits	5 Year Trend
yonik	135	1447	
(Manager)	540	890	III.III. III.II. II. JUIIII.
muir	157	599	S year commit count

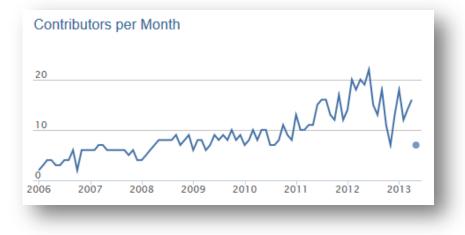


Name	12 Month Commits	All Time Commits	5 Year Trend
(founder, developer)	0	2478	5 year commit count
Shay Banon	642	1504	5 year commit count
Martijn van Groningen	165	165	5 year commit count
Igor Motov	119	152	5 year commit count

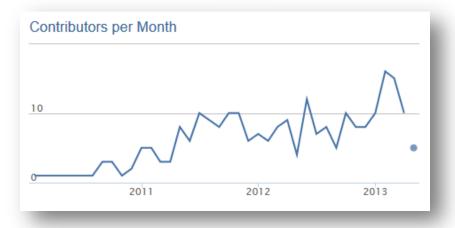


## **Active Contributors**





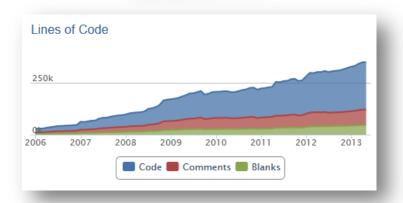


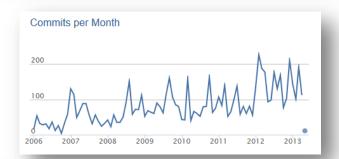




## **Lines of Code**

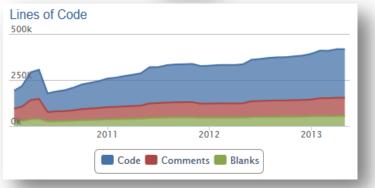


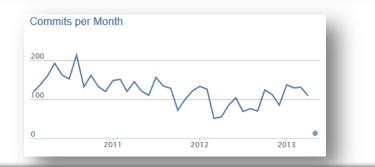












## **The Mailing Lists**



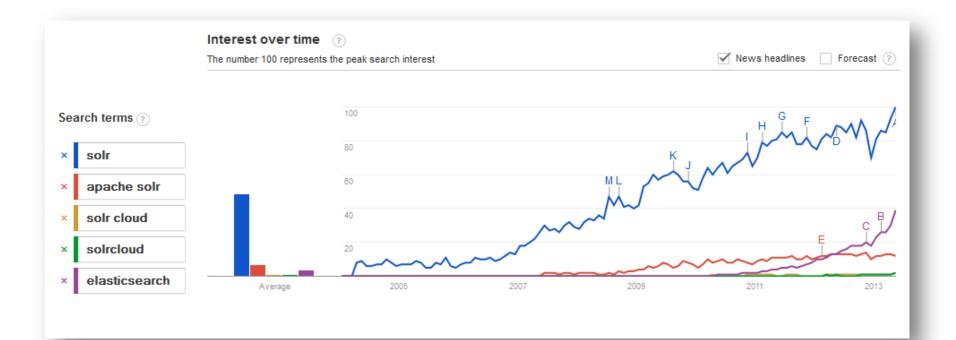








## Interest over time

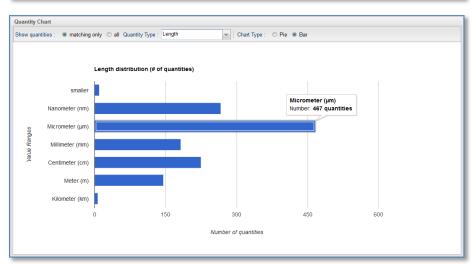




## **Case study - quantalyze**



compile quantatyze query	🏓 Search Results (I) 🔮 Quantity Chart 🔮 Filter Chart					
Query Fitter (-)	✓ Result List				Export 0	
Full test query	Publication No. Trie	Quant * Assignee	Publ. Date	Priority No.	Priority Date	
mLip*	US658301 (B1) Linds	1.650 Celltech R & D Limited		GE9807727	1998-04-08	
	US6/12681 (B1) Method for extracting compounds of furan lipids and polyhydrosolated fatty alcohols of avocado, composition based on said compoun	86 Pharmascience	2003-06-24	FR99 11846	1959-09-22	
Quartity query	US6682957 (B1) Uppryoenase poterns and ophroutlectides encoding the same	19 Lexicon Genetics Incorpo				
latus formatis, value, value+colecance, lower, upper, lower,	organization (nu) in the design and the design and a second meritance.	To concer official a nearly	and Lour Veres			
Combine restrictions using:						
Section Value Unit Context						
Vhole Docume -10. 200 Celsius (°C) 🕬 contexts 🔀	✓ Document			U\$6583301 (B	1) Lipide	
	(D12) N-aninoethyl-O+Butylcarbamate	Stow quartities : 0 ms	Show quantities : O matching only # all			
	BOC-04 (15.4.4.1.506 mmel) was added to a stirred mixture of ethylenedismic (13.4.ml. 0.2 mmel) and treethylenine (21.ml. 0.2 mmel) at store where a on hold have a spon and let ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where an on-hold have a spon and let ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where an on-hold have a spon and let ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where an on-hold have a spon and leth ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where an on-hold have a spon and leth ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethyl accutate was added and the product ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mmel) at store where a spon and leth ownright. Ethylenine (21.ml. 0.2 mm				Field	
	washed with brine, dired and evaporated to dryness to yield the title compound which analys- Paund Ese, MH+ 951, MI (COCI) 5: 21 (H, to COH), 3 19 (2H, et COCINC(2), 2 31 (H) Million (H, et M), Million (H), Mill	0. nitrile (500 ml) at mom	temperature for two	days. The a	Desc	
		methane (100 ml) at =	8" C. was carefully	added	Desc	
	To D11 (2.041.g. 17.4 mmol) in dry dichloromethane (50 ml) under argon was added triathylamine (1.761.g. 17.4 mmol). EDC (5.171.g. 17.4 mmol) and N-hydronyuccinimide (2.002.g. 17.4 mmol). The solution was left for three hours, during which time the formation of the slower moving MIS active enter was	the temperature at -7	C The reaction w	no st.	Desc	
	followed by silica tic (r.f. 0.25, 35% ethyl acetate in hexane). Once complete ester formation was achieved, D12 (3.067 g, 15.1 mmd) was added in	an allowed by speed house	m temperature Wa	ter (100 ml) sans	Desc	
	dichloromethane (2014) and the reaction left states. Purification on silica (2014 edit) acetate in hexane) yielded the title compound (2005 g. 445%), silica 0 0 15 (2014 edit) acetate in hexane). C32H52K2O5 requires 554.47. Found ES+: MH+ 555.40. MHa+ 557.41. 8H (CDCI)) 0 7.0.9 (12H. m. CHitle). 0.0.1 8 (2BH)	L br mil under angen at "0" (			Desc	
	m, CHMa, CH2), 1.41 (9H, s. C(Me)3), 2.15 (2H, 6H, CH2CONH), 2.27 (2H, m, CH2CO2Me), 3.15-3.35 (6H, m, CH2NH), 3.63 (2H, s. OMe), 5.33 (1H, t. NHCC), 6.65 (1H, t. NHCC), 8C (CDC33), 19.1, 19.5, 24.2, 27.3, 28.2, 31.7, 32.2, 32.6, 34.3, 37.0, 295 CH2, CH3A, CDCH33, 48.2, 48.4 (2C, CH2N), 51.2 (1C, OMe), 6.65 (1H, t. NHCC), 8C (CDC33), 19.1, 19.5, 24.2, 27.3, 28.2, 31.7, 32.2, 32.6, 34.3, 37.0, 295 CH2, CH3A, CDCH33, 48.2, 48.4 (2C, CH2N), 51.2 (1C, OMe), 6.65 (1H, t. NHCC), 8C (CDC33), 19.1, 19.5, 24.2, 27.3, 28.2, 31.7, 32.2, 32.6, 34.3, 37.0, 295 CH2, CH3A, CDCH33, 48.2, 48.4 (2C, CH2N), 51.2 (1C, OMe), 37.0 (2H, t. NHCC), 38.0 (2H, t		for 1/2 hour and t		Deac	
	79.2 (1C, C(Me)3), 156.8 (1C, 194C02), 174.2, 174.4 (2C, CO2Me, CONH).	n was maintained at 0*			Desc	
	(D14) N=6 8utylanycarbonylaminosthyl) 4,0 13,17 tetramethyl-1,20 dodecanamolic Acid To D13 (6286 6, 7722 mmd) in methanol (35 mi) was added a suscension/solution of lithrum hydroxide (1238 6, 77.22 mmd) in system (35 mi). This was	s a low melting (No 21	Chabite solid Cit	2144	Desc	
	stirred for 2 hours then poured into 100% citric acid solution and extracted with ethyl acetate. These organic fractions were combaned, washed with water and th				Desc	
	solvert removed to yield the title compound. With dichloromethane methanol water 6 1.1 on silica the tic of D13 and D14 have Rf of 0.35 and 0.05 respectively C31H600205 requires 540.65. Found ES+ MHe 541.48. MNe+ 563.45. AH (CDC3) 0.05-1.0 (12H, m. Me1.1.0-18./28H, tr.m. CH2. CH1.1.46./3H, a. C0MeX3				Desc	
	2 20 (2H. m. CH2CONH), 2 36 (2H. m. CH2CO2), 3 15-3 45 (4H. m. CH2N), 5 16 (1H. I. NHCO2), 6 55 (1H. I. NHCO3), 6C (CDC3) 19-3, 19-7, 24-2, 27-2, 28-3				Desc	
Add quantity restriction	31 8, 32 3, 32 6, 34 4, 36 9 (25C, CH2, CHMe, C(CH3)), 40 0, 40 6 (2C, CH2N), 79 6 (1C, C(Me)3), 157 9 (1C, NHCO2), 174 5 (1C, CONH), 178 5 (1C, CO2H) (D15) N-Aminoethyl-4, 8, 13, 17-stramethyl-1, 20-dodecaramoic Acid Trifucoracetate Salt		1M NaDH (50 ml) at 50° C BOC antrodide (2 33		Desc	
And dramay insurroun	D14 was taken up into 96% TFA and left for 10 menutes. The solvent was removed to give the amino acid as a viscous oil which after silica chromatography (1		e (2 3) g. 74%). Mp 72:76° C. C17H3NO4 requires			
Search	mathanol in dichloromethane+0 1% acetic acid) gave the title compound as a colourless solidigum, silica tito r1 0.25, ninhydtin +w (6.1 dichloromethane methanol). The yield for the two steps 014 to 015 was 2004 a. 80% C26H52N2O3 requires 440.40. Found ES+: MH+ 441.36, BH (C03OD)		d (2 33 g. 50%) Mp 85-87° C. C25H5R(205 requires		Desc.	
	0.95-1.1 (12H, m. Me), 1.1.1.9 (2BH, Sr m. CH, CH2), 2.38 (4H, m. CH2CO), 3.29 (2H, t. CH2NH3+), 3.60 (2H, t. CH2NHCO), 8C (CD3OD) 20.1, 20.5, 22.0, 25.	E	acid (2.12 a) Mo 66-67 C. C29H68QCO requires			
Hint If you spot a wrongly identified measurement value/internal, you can	28 6, 33 6, 33 8, 34 0, 35 0, 38 4, 38 5 (22C, CH, CH2, Me), 49 9 (2C, CH2H), 177 9 (1C, CORH), 179 4 (1C, CO2H), C24 Ameracid Centaring Mid-chain Amide				Desc	
click on it with the right mouse button to notify us. Conversely, mark a	(D16) 12-8 Butylonycartionylamino)lobeicaroic Acid BochH-iiCH2(11-CD2H	pension was left at -10				
section of the text containing a physical quantity that has been missed by our analysis engine and also use the right mouse button to open	12-Amenododecanoic acid (215 g. 10 mmol) was dissolved in 1M NaOH (50 ml) at 50° C. BOC anhydride (2.33 g. 10 mmol) was added to the reaction while				Desc.	
the feedback form. Thank you for helping up improving our quantalyze service.	was stimed for 30 mins. The reaction was poured into stimed 100% citric acid, 100 ms) and the white solid filtered off. The solid was washed with citric acid wate and dried in vacuo. The product was dissolved in effect and filtered before evaporating to dryness. The title corroound was crystalised from hexane (2008), 744				Desc.	
many you to needing on independ on draugstyce revice.	Mp 72-76* C 017H03N04 requires C 64 73% H 10 55% N 4 44% Found C 64 75% H 10 58% N 4 45% C24H89N02 requires 315 Found ES+ 5	Alt			Desc.	
	316 AH (CDC0) 3 (8) (2H + CH0M0) 2 33 (2H + CH0C02H) 16 (2H m CH0CH2H2) 144 (3H mm CH0CH2C02H4Brc) 126 (3H m CH0)	ation was heated at #2				



patents.quantalyze.com

- Runs in your browser
- Filter and keyword search
- Physical quantity search
- Interval search
- Print view

#### Visualization:

- Physical quantity distributions
- Cross-tabulations (e.g. concepts vs. quantity type)
- Different chart types



## Case study - StumbleUpon



## Create a world-class customer experience

- A "Stumble" provides real-time recommendations to 30 million customers per day
- Intelligent search is key to providing fast and more informed recommendations
- Update your searches immediately with newly posted content

#### **Develop and scale easily**

- Build in intelligent search to scale with millions of users and interactions
- Take advantage of powerful and flexible APIs for easy data integration
- Use easy to use but powerful solutions for your big data search and analytics needs



## Strengths of open source search

- Best practice segmented index (like Google, Fast)
- Scalability
- Best practice, flexible ranking (term/field/doc boosts, function queries, custom scoring...)
- Best overall query performance and complete query capabilities (unlimited Boolean operations, wildcards, findsimilar, synonyms, spell-check...)
- Multilingual, query filters, geo search, memory mapped indexes, near real-time search, advanced proximity operators...
- Rapid innovation
- Extensible architecture, complete control (open source)
- No license fees (open source)



## Weaknesses of open source search

- Those typical of open source
  - No formal support
  - Limited access to training, consulting
  - Lack of stringent integrated QA
  - Speed of development and open source environment too complex for some (e.g., what version should I download? What patches? GUI?)
- Others
  - Lucene/Solr/Elasicsearch development has tended to focus on core capabilities, so missing certain features for enterprise search (e.g., connectors, security, alerts, advanced query operations)



## Addressing open source weaknesses

- Community
  - Community has a wealth of information on web sites, wikis and mailing lists
  - Community members usually respond quickly to questions
- Consultants
  - May be especially helpful for systems integration or addressing gaps
- Commercialization
  - Companies commercializing open source provide commercial support, certified versions, training and consulting
- Internal resources



## Product strengths of commercial competitors

- Well established players tend to be full-featured
- Some organizations have focused on a particular application or domain (e.g., ecommerce, publishing, legal, help desk)
- Some competitors have focused on appliance



## Weaknesses of top commercial competitors

- Usually expensive, especially at scale
- Platform or portability limitations
- Limited transparency
- Limited flexibility, especially for other than intended application or domain
- Limited customization, especially for appliance-like products
- Sometimes limited scalability
- Technical debt and/or lack of rapid innovation
- Customers are dependent on the company's continued business success



## **Competitive landscape**

- Last years commercial companies have felt increasing competition from Lucene/Solr/Elasticsearch because of the combination of its capability and price
- Some competitors have responded with diversification
- Some have been acquired
- Need for good, affordable, flexible search remains



### Questions









## Credits



## sematext

