

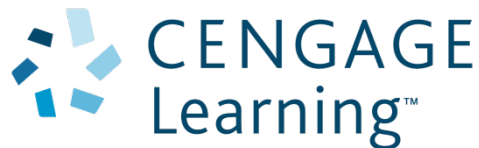
APACHE
LUCENE
EUROCON



Relevance Improvements at Cengage

Ivan Provalov, Cengage Learning
ivan.provalov@cengage.com, 10/19/2011

Presented by



Outline

- Relevance improvements at Cengage Learning
- My background
- Cengage Learning
- Relevance issues
- Relevance tuning methodology
- Relevance tuning for English content
- Non-English content search
- Conclusions

My Background

- Ivan Provalov
 - Information Architect, Search
 - New Media: Technology and Development
 - Cengage Learning
- Background
 - Software development, architecture
 - Agile development
 - Information Retrieval, NLP
- Michigan [IR Meetup](#)
 - Interested in presenting IR topic online?

Cengage Learning



- \$2B annual revenue
- One of the world's largest textbook publishers and provider of research databases to libraries
- Academic and Professional
 - **Brands: Gale, Brooks/Cole, Course Technology, Delmar, Heinle, South-Western, and Wadsworth**
 - **National Geographic School Publishing**
- International locations
 - **US, Mexico, Brazil, UK, Spain, India, Singapore, Australia, China**
- Aggregated periodical products
 - **15,000 publications**

Search Platform Profile

- Lucene 3.0.2
- Supporting approximately 100 products
- Content rights management
- Custom products
- 210 million documents
- 150 index fields per shard
- 21 million terms per shard's full text index
- 250 queries per second

Subjects

- Arts
- Biography
- Business
- Education
- Environment
- General Reference
- History
- Information and Publishing
- Law
- Library Science
- Literature
- Medicine
- Multicultural Studies
- Nation and World
- Religion
- Science
- Social Science
- Technology

Arts

[View All](#)



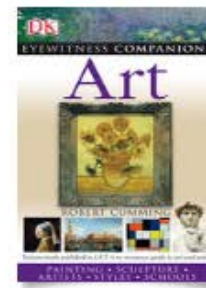
American Folk Songs: A Regional Encyclopedia
2008



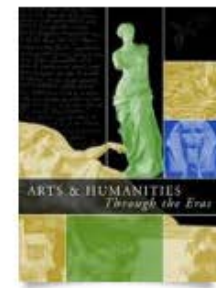
The American Repertory
2005



Architecture of Italy
2008



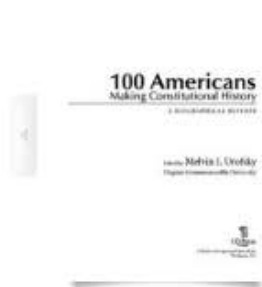
Art, 1st American ed.
2005



Arts and Humanities
2005

Biography

[View All](#)



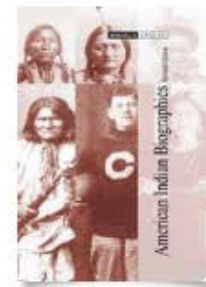
100 Americans Making Constitutional History
2004



African Americans in Science
2008



Al Capone: A Biography
2003

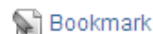


American Indian Biographies, Rev. ed.
2005



American Men & Women of Science
2007

Search Results



Search results: Basic Search= Stock prices x LIMITER: Full Text Documents= Y x

Everything

[Viewpoints \(1\)](#)

[Images \(14\)](#)

[News \(21,995\)](#)

[Magazines \(216\)](#)

[Academic Journals \(44\)](#)

[Audio \(168\)](#)

[Videos \(14\)](#)

[Websites \(36\)](#)

Search within results



[Viewpoints for Stock prices](#)



Save

Back Dating Stock Options Undermines Trust in Business

Corporate Corruption, 2007

"Back Dating Stock Options Undermines Trust in Business" by Jeff Brown. *Corporate Corruption* Susan Hunnicutt, Ed. At Issue Series. Greenhaven Press, 2007. Jeff Brown, "The Problem With Back-Dating Stock Options," ...

[Images for Stock prices](#)



Save

Save

Save

Save

Save

Save

Save

Save

Save

Save

[News for Stock prices](#)



Save

Stocks register marginal gains

The Manila Times (Manila, Philippines), September 30, 2011

Byline: Krista Angela M. Montealegre Sept. 30--PHILIPPINE share prices posted marginal gains on Thursday as investors hunted for bargains despite the weakness in overseas markets on renewed Greek debt woes. At the ...



Save

Bralirwa Share Price Hits Record High

Africa News Service, September 29, 2011

Sep 29, 2011 (The New Times/All Africa Global Media via COMTEX) -- Shares of Rwanda's leading brewery, Bralirwa, rose by 1.7 per cent during yesterday's trading at the Rwanda Stock Exchange (RSE). The brewery

Mostly Search Guys (MSG)



Relevance Issues

- Ranking
 - Recency, publication ranking
 - Document length
 - Unbalanced shards
- Zero results queries
- Content quality
 - Near duplicate documents
 - Digitization
- Language
 - Tokenization
 - Detection

Relevance Tuning Methodology

- TREC
- Metrics: MAP, MRR, Recall
- Internal studies of top results
- Outsource to universities (WSU)
- Relevance feedback web application
- Usage logs review

Relevance Evaluation Form

Result:

Potential Energy

POTENTIAL ENERGY

Potential energy is the energy that something has because of its position or because of the arrangement of its parts. A baseball in flight has potential energy because of the arrangement of the atoms in the molecule.

A diver standing on a platform above the water has potential energy because of a capacity for doing work by jumping off. The higher the diver is above the water, the more potential energy he or she has. The potential energy of the diver is measured in joules (J) and is given by $U = mgh$, where m is the mass in kilograms (kg), g is the acceleration due to the gravitational force (9.80 meters per second squared), and h is the height in meters. A 60 kg diver has 2,350 J of gravitational potential energy.

Any type of potential energy is associated with a force and involves some favorable position. A compressed spring in a toy gun does work on a projectile when it is released. The force involved is that of the spring, and the favorable position is the amount of compression. Other examples of potential energy are the energy of a pole, compressing the suspension springs of an automobile, and stretching a rubber band.

A loss in potential energy by a system is accompanied by a gain of energy in some other form. All the potential energy of a mass held above a spring is converted to kinetic energy as it pushes against the spring, but the spring gains potential energy as a result of being compressed. Water atop a dam in a hydroelectric power plant has potential energy. When the water impinges on the blades of a water turbine, the rotor of the turbine gains rotational kinetic energy at the expense of the potential energy of the water. Molecules have potential energy associated with electric forces that bind the atoms together. In a chemical reaction liberating energy—heat, for example—the potential energy is accompanied by an increase in other forms of energy.

Protons and neutrons in the nucleus of an atom have potential energy associated with nuclear forces. In a nuclear reaction liberating energy, the nuclear energy is accompanied by an increase in other forms of energy. A single nuclear reaction in a nuclear power plant liberates nearly 10 million times the energy of a chemical reaction.

Joseph Priest

See also ; ; ; .

BIBLIOGRAPHY

Hobson, A. (1995). *Physics: Concepts and Connections*. Englewood Cliffs, NJ: Prentice-Hall.

Priest, J. (2000). *Energy: Principles, Problems, Alternatives*, 5th ed. Dubuque, IA: Kendall/Hunt Publishing Co.

Serway, R. A. (1998). *Principles of Physics*, 2nd ed. Forth Worth, TX: Saunders College Publishing.

Query: "energy"

Select a score:

- 0: no document available
 1: not relevant
 2: somewhat relevant
 3: very relevant

[View Score Definitions](#)

You may enter a comment below:

[Add Evaluation](#)

Relevance Report by Query Type

Query Type	Option A	Option B	Option C
long tail - misspelled	1.225	1.271	2.269
long tail - multi-term	1.502	1.885	1.929
top 100 - frequent expression	2.574	2.013	2.595
top 100 - person	2.689	1.904	2.778
top 100 - place	1.863	1.315	2.438
top 100 - single-term	2.753	1.775	2.792
top 100 - work	2.408	2.198	2.514
Grand Total	2.377	1.822	2.481

Usage Data Reporting

USAGE LOG REPORTS OPTIONS

Report Group: Report Date:

Report Type: Select Date is : Apr 26, 2011

REPORTS LIST

- session_duration_summary_report.html
- session_query_summary_report.html
- session_retrieval_summary_report.html

100 Most Used GVRL Queries Report

Query	Query Usage
KE (hamlet)	183
KE (absolutism)	131

Relevance Improvements

- Pre-processing
 - Search assist
- Search time
 - Term query expansions
 - Phrase query expansions
 - Score modifications
- Post-processing
 - Results clustering

Search Assist

- Search terms recommendation
- Predictive spelling correction for queries
- Commonly occurring phrases in the content
- Over 6 million terms in keyword dictionary
- Limited by relevant content sets

Term Query Expansions

- Stemming
 - Porter stemmer
 - Dynamic
 - Stem families
{water -> waters, watering, watered}
- Spelling
 - Dictionary size – 125K
 - Skipping when use other phrase-based expansions
{pharmaceuticall -> pharmaceutical}
- Pseudo relevance feedback
 - More like this

Phrase Query Expansions

- Related subject synonyms
 - **Controlled vocabulary**
 - **Machine aided indexing, manual indexing**
{death penalty -> "capital punishment"}
- Phrase extractions
 - **Based on POS pattern (e.g. JJ_NN_NNS)**
{afro-american traditions, religion and beliefs -> afro american traditions}

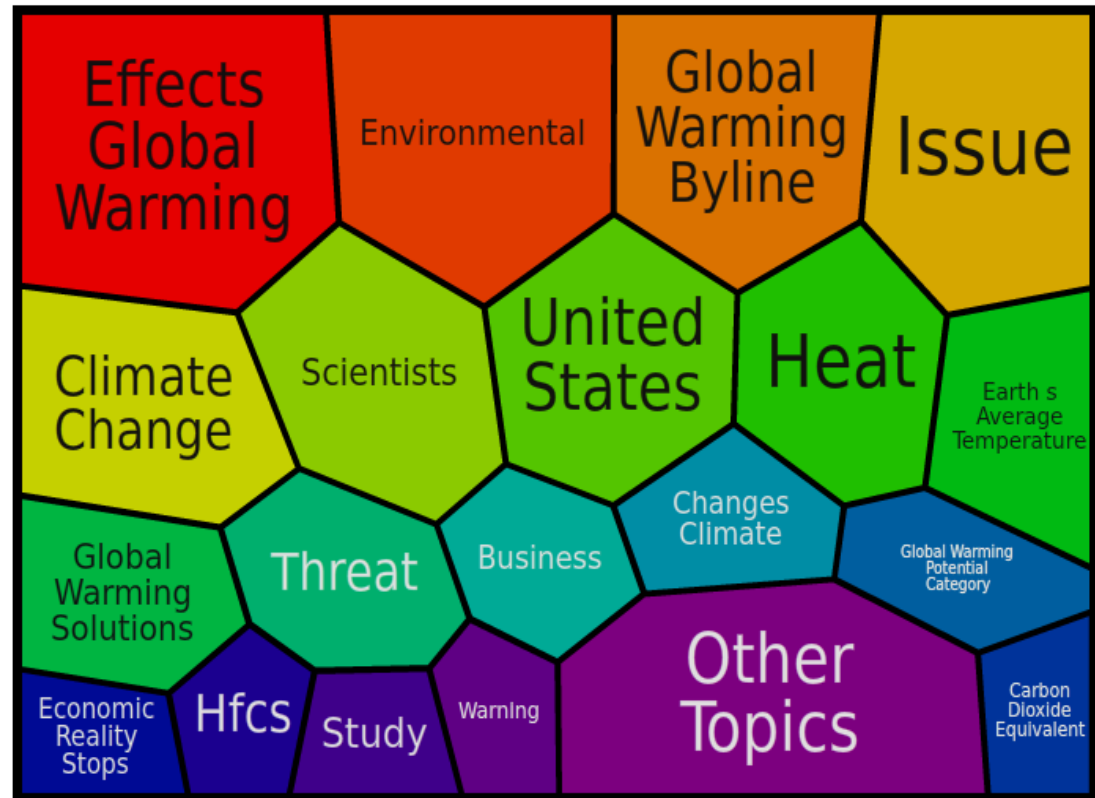
Score Modifications

- Recency boosting
 - Function Query
 - Tunings – dates range, boost
- Publication boosting
 - Function Query
 - Publication list with importance ranks
 - Clickstream mining
 - Peer reviewed

Results Clustering

- Subject index
- Dynamic: Carrot2

[global warming \(545\)](#)
[Climate change \(336\)](#)
[Human-environment in... \(126\)](#)
[Greenhouse gases \(120\)](#)
[Environmental policy \(89\)](#)
[Emissions \(Pollution... \(81\)](#)
[Greenhouse effect \(72\)](#)
[carbon dioxide \(69\)](#)
[Environmental protec... \(66\)](#)
[Environmental degrad... \(65\)](#)
[Climate \(52\)](#)
[Atmospheric carbon d... \(42\)](#)
[Protocol to the Fram... \(41\)](#)
[Air pollution contro... \(41\)](#)
[Environmental resear... \(39\)](#)
[Surface-ice melting \(37\)](#)
[Environmentalism \(36\)](#)
[sea level \(33\)](#)
[Intergovernmental Pa... \(30\)](#)
[Water cycle \(28\)](#)
[sustainable developm... \(28\)](#)
[Ice sheets \(28\)](#)
[Solar radiation \(28\)](#)
[environmental manage... \(28\)](#)
[Ice ages \(28\)](#)



Spanish

- Datasets
- Stemmer
- Stopwords
- Encoding

	Lucene Standard Analyzer	Lucene Spanish Analyzer
MAP	0.326	0.401
MRR	0.781	0.8
Recall	0.649	0.716

Arabic

- Datasets
- Letter tokenizer
- Normalization filter
- Stemmer

	Lucene Standard Analyzer	Lucene Arabic Analyzer
MAP	0.192	0.296
MRR	0.516	0.615
Recall	0.619	0.77

Chinese

- 60k+ characters with meaning
- No whitespace
- Sorting
- Encoding
- Hit highlighting

Analyzer	MAP
ChineseAnalyzer	0.384
CJKAnalyzer	0.416
SmartChineseAnalyzer	0.412
IKAnalyzer	0.409
Paoding	0.444

Conclusions

- Lucene relevance ranking
- Simple techniques (stemming, PRF, recency)
- Language specific analyzers
- TREC collections
- Relevancy priority in search system development

Contact

- Ivan Provalov
 - Ivan.Provalov@cengage.com
 - <http://www.linkedin.com/in/provalov>
- IR Meetup
 - <http://www.meetup.com/Michigan-Information-Retrieval-Enthusiasts-Group>

Acknowledgements

- *Cengage Learning*: J. Zhou, P. Tunney, J. Nader, D. Koszewnik, J. McKinley, A. Cabansag, P. Pfeiffer, E. Kiel, D. May, B. Grunow, M. Green
- *Lucid Imagination*: R. Muir
- *University of Michigan*: B. King
- *Wayne State University*: B. Hawashin, Y. Wan, H. Anghelescu
- *Michigan State University*: B. Katt
- *HTC*: R. Laungani, K. Krishnamurthy, R. Vetrivelu
- *University of Mich. Library*: T. Burton-West

References

- Relevance [http://en.wikipedia.org/wiki/Relevance_\(information_retrieval\)](http://en.wikipedia.org/wiki/Relevance_(information_retrieval))
- Lucene in Action, 2nd Edition, Michael McCandless, Erik Hatcher, and Otis Gospodnetić
- Introduction to IR <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>
- TREC <http://trec.nist.gov/tracks.html>
- IBM TREC 2007 <http://trec.nist.gov/pubs/trec16/papers/ibm-haifa.mq.final.pdf>
- L. Larkey, et al, Light Stemming for Arabic Information Retrieval. Kluwer/Springer's series on Text, Speech, and Language Technology. IR-422, 2005.
- Cengage Learning at the TREC 2010 Session Track, B. King, I. Provalov, <http://trec.nist.gov/pubs/trec19/papers/gale-cengage.rev.SESSION.pdf>