DOMAIN-SPECIFIC KNOWLEDGE-GRAPH CONSTRUCTION AND APPLICATIONS (WITH ALMOST NO SUPERVISION)

Maya Ramanath

IIT-Delhi

What are Knowledge-bases?

Subject	Predicate	Object
Albert_Einstein	bornIn	Germany
Albert_Einstein	typeOf	Theoretical_Physicist
Albert_Einstein	wonAward	Nobel_prize_in_Physics
Titanic	directedBy	James_Cameron
Titanic	genre	Drama
Topological_sorting	typeOf	Graph_algorithm
Group_Testing	applicationOf	Coding_theory
Gigabit_ethernet	highSpeedFormOf	Ethernet
Iron_python	adaptationOf	Python
Barack_Obama	supports	Military_Action_in_Libya
Donald_Trump	opposes	Illegal_immigration
UK	supports	Brexit
Scotland	opposes	Brexit
Pseudonomas	foundIn	Sea
Pseudonomas	actsOn	Acetone
Hydrocarbon_degradation	occurIn	Sea
Hydrocarbon_degradation	carriedOn	Acetone

Facts about people, movies

Concepts in Computer Science and their relationship with each other

Opinions about important(?) political topics

Facts about entities in bio-chemical engineering

Turn the Web…

news, books, entertainment, scientific content, advertisements, photos, videos, sound, social networks, blogs, tweets, opinions, comments, spam, junk...

…into a Knowledge-base

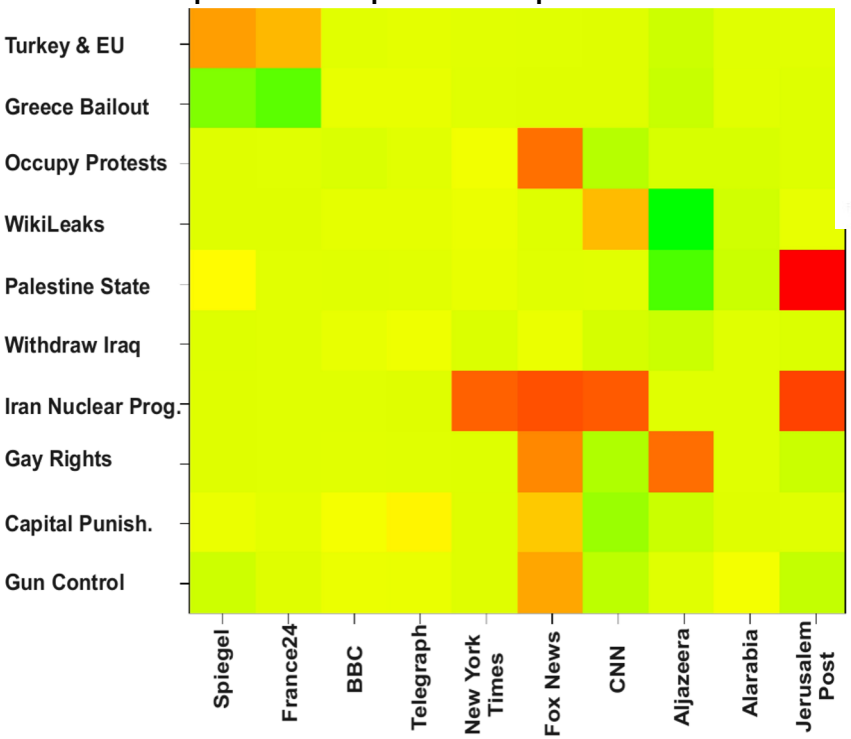
well organized, crisp, machine-readable, browseable, searchable, self-maintaining encyclopedia human knowledge.

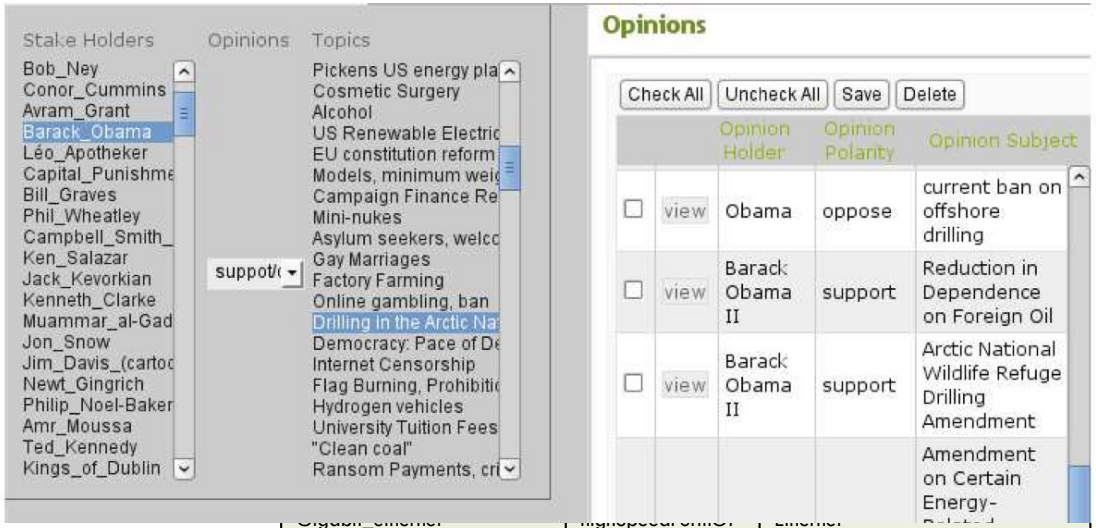
Outline

- Introduction
- Applications
- Construction
- Conclusion

Applications of KBs

1. Opinion Base [WSDM 2012, CIKM 2012, PLEAD@CIKM
   - "Explore the space of opinions"





Iron_python	adaptationOf	Python
Barack_Obama	supports	Military_Action_in_Libya
Donald_Trump	opposes	Illegal_immigration
UK	supports	Brexit
Scotland	opposes	Brexit
Pseudonomas	foundIn	Sea
Pseudonomas	actsOn	Acetone
Hydrocarbon_degradation	occurIn	Sea
Hydrocarbon_degradation	carriedOn	Acetone

George W. Bush & Abortion

(+−)
(−+)
arah Palin⟩ ⟨+⟩ ⟨Ge
[⟨Sarah Palin⟩ ⟨-⟩

(++)
(−−)
Joe Biden⟩ ⟨+⟩ ⟨George W. Bush⟩]
⟨Joe Biden⟩ ⟨+⟩ ⟨Abortion rights⟩]

(−+)
[⟨Barack Obama⟩ ⟨+⟩ ⟨NDAA⟩]
[⟨Barack Obama⟩ ⟨-⟩ ⟨Guantanamo Bay⟩]

Applications of KBs

1. Opinion Base
   - "Explore the space of opinions"
2. TeKnowBase (Academic search)
   - How can I search for research articles discussing algorithms for autoencoders?

Subject	Predicate	Object
Albert_Einstein	bornIn	Germany
Albert_Einstein	typeOf	Theoretical_Physicist
Albert_Einstein	wonAward	Nobel_prize_in_Physics
Titanic	directedBy	James_Cameron
Titanic	genre	Drama
Topological_sorting	typeOf	Graph_algorithm
Group_Testing	applicationOf	Coding_theory
Gigabit_ethernet	highSpeedFormOf	Ethernet
Iron_python	adaptationOf	Python
Barack_Obama	supports	Military_Action_in_Libya
Donald_Trump	opposes	Illegal_immigration
UK	supports	Brexit
Scotland	opposes	Brexit
Pseudonomas	foundIn	Sea
Pseudonomas	actsOn	Acetone
Hydrocarbon_degradation	occurIn	Sea
Hydrocarbon_degradation	carriedOn	Acetone

"Autoencoder"

Semantic Scholar | All Fields | autoencoder

About 34,000 results | Last Five Years | Lit Reviews | Has PDF | Has Video | More Filters

Extracting and composing robust features with denoising autoencoders
Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol • ICML • 2008
Previous work has shown that the difficulties in learning deep generative or discriminative models can be overcome by an initial unsupervised learning step that maps inputs to useful intermediate... (More)
299 478 1 View on ACM View Slides Cite Save

Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection
Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, Christopher D. Manning • NIPS • 2011
Paraphrase detection is the task of examining two sentences and determining whether they have the same meaning. In order to obtain high accuracy on this task, thorough syntactic and semantic analysis... (More)
81 74 View PDF Cite Save

Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions
Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, Christopher D. Manning • EMNLP • 2011
We introduce a novel machine learning framework based on recursive **autoencoders** for sentence-level prediction of sentiment label distributions. Our method learns vector space representations for... (More)
104 118 View PDF Cite Save

Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion
Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol • J. Mach. Learn. Res. • 2010
We explore an original strategy for building deep networks, based on stacking layers of denoising **autoencoders** which are trained locally to denoise corrupted versions of their inputs. The resulting... (More)
370 551 View PDF Cite Save

Semantic Scholar | All Fields | autoencoder algorithm

About 28,400 results | Last Five Years | Lit Reviews | Has PDF | Has Video | More Filters

Extracting and composing robust features with denoising autoencoders
Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol • ICML • 2008
Previous work has shown that the difficulties in learning deep generative or discriminative models can be overcome an initial unsupervised learning step that maps inputs to useful intermediate... (More)
299 478 1 View on ACM View Slides Cite Save

Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions
Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, Christopher D. Manning • EMNLP • 2011
We introduce a novel machine learning framework based on recursive **autoencoders** for sentence-level prediction o sentiment label distributions. Our method learns vector space representations for... (More)
104 118 View PDF Cite Save

Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network wi Local Denoising Criterion
Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol • J. Mach. Learn. Res. • 2010
We explore an original strategy for building deep networks, based on stacking layers of denoising **autoencoders** wh are trained locally to denoise corrupted versions of their inputs. The resulting... (More)
370 551 View PDF Cite Save

Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection
Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, Christopher D. Manning • NIPS • 2011
Paraphrase detection is the task of examining two sentences and determining whether they have the same meanin order to obtain high accuracy on this task, thorough syntactic and semantic analysis... (More)
81 74 View PDF Cite Save

Query + aspect: A few results

Query	*Application*	*Algorithm*	*Implementation*
`cryptography`	Cryptographic Protection of Computer-Based Data Files	Cryptographic algorithm on multicore processor: A review	Design of a Reconfigurable Hardware for Efficient Implementation of Secret Key and Public Key Cryptography
`autoencoder`	Exploring autoencoders for unsupervised feature selection	Training Stacked Denoising Autoencoders for Representation Learning	Parallelizing the Sparse Autoencoder
`neural_network`	Application of Particle Swarm Optimization and RBF Neural Network in Fault Diagnosis of Analogue Circuits	Evolutionary Neural Network Based on New Ant Colony Algorithm	Design of FPGA based general purpose neural network
`hashing`	Sampling Based N-Hash Algorithm for Searching Frequent Itemset	An Alternative Analysis of the Open Hashing Algorithm	Low Power And Area Efficiency of SHA-1

Aspect-based Search [ECIR 2020]

Topological_sorting	typeOf	Graph_algorithm
Group_Testing	**applicationOf**	**Coding_theory**
Gigabit_ethernet	highSpeedFormOf	Ethernet
Iron_python	adaptationOf	Python
Feature_selection	**applicationOf**	**Autoencoders**
Heap_sort	**algorithmFor**	**Sorting**

- A LM-based method for aspect-aware search

$$MM(w) = \lambda P(w|a) + (1-\lambda) P(w|q,a)$$

Prob. of term, given aspect

"Application"

Prob. of term, given both query and aspect

"Autoencoder"+

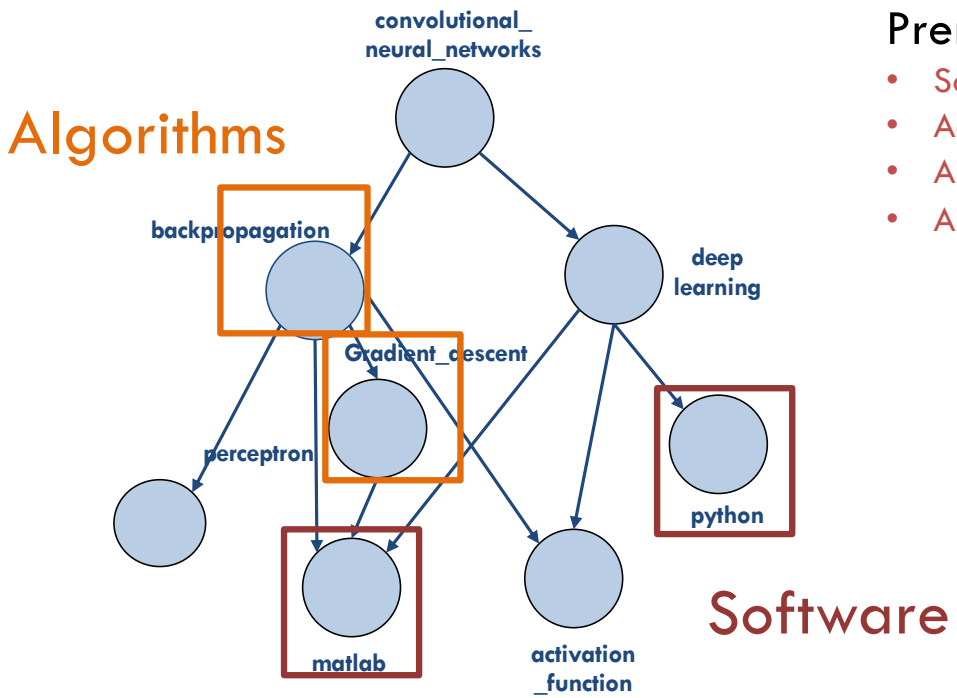
Approach	Algorithm			Application			Implementation		
	DCG@5	P@5	P@1	DCG@5	P@5	P@1	DCG@5	P@5	P@1
MM	**6.27**	**0.70**	**0.75**	**2.64**	**0.45**	**0.47**	**2.33**	**0.44**	0.40
QL+query	2.69	0.3	0.33	1.42	0.25	0.22	1.05	0.16	0.23
QL+query+aspect	5.03	0.56	0.59	2.38	0.41	0.35	1.92	0.30	*0.43*
QL+query+aspect+QE	*5.12*	*0.58*	*0.61*	*2.5*	*0.43*	*0.41*	*2.29*	*0.37*	**0.49**

Applications of KBs

1. Opinion Base
   - "Explore the space of opinions"
2. TeKnowBase (Academic search)
   - How can I search for research articles discussing algorithms for autoencoders?
   - What are the pre-requisites to learn about convolutional neural networks?

Subject	Predicate	Object
Albert_Einstein	bornIn	Germany
Albert_Einstein	typeOf	Theoretical_Physicist
Albert_Einstein	wonAward	Nobel_prize_in_Physics
Titanic	directedBy	James_Cameron
Titanic	genre	Drama
Topological_sorting	typeOf	Graph_algorithm
Group_Testing	applicationOf	Coding_theory
Gigabit_ethernet	highSpeedFormOf	Ethernet
Iron_python	adaptationOf	Python
Barack_Obama	supports	Military_Action_in_Libya
Donald_Trump	opposes	Illegal_immigration
UK	supports	Brexit
Scotland	opposes	Brexit
Pseudonomas	foundIn	Sea
Pseudonomas	actsOn	Acetone
Hydrocarbon_degradation	occurIn	Sea
Hydrocarbon_degradation	carriedOn	Acetone

Pre-requisites for "convolutional_neural_networks"



Prerequisites for which aspect?

- Software: matlab, python
- Activation functions: sigmoid, relu, tanh, softmax
- Applications: face perception, object identification
- Algorithms: backpropagation, optimization, gradient descent

Subject	Predicate	Object
Topological_sorting	typeOf	Graph_algorithm
Group_Testing	typeOf	Permutation
Gigabit_ethernet	typeOf	Permutation
Iron_python	typeOf	Permutation
Bubble_sort	typeOf	Permutation
Sorting	typeOf	Permutation
Sorting	typeOf	Computation
Feature_selection	applicationOf	Autoencoders
Bubble_sort	hasComplexity	$O(n^2)$
Backpropagation	algorithmFor	Neural_network
CNN	typeOf	Neural_network

Techniques	Precision
PreFace	0.76
QDMKB + RefD	0.636
RefD + TKB	0.68

PreFace [ISWC 2020, ECIR 2021]

Applications of KBs

1. Opinion Base
   - "Explore the space of opinions"
2. TeKnowBase (Academic search)
   - How can I search for research articles discussing algorithms for autoencoders?
   - What are the pre-requisites to learn about convolutional neural networks?
3. Biochemical KB
   - "Determine if a microbe is aerobic or anaerobic"

Subject	Predicate	Object
Albert_Einstein	bornIn	Germany
Albert_Einstein	typeOf	Theoretical_Physicist
Albert_Einstein	wonAward	Nobel_prize_in_Physics
Titanic	directedBy	James_Cameron
Titanic	genre	Drama
Topological_sorting	typeOf	Graph_algorithm
Group_Testing	applicationOf	Coding_theory
Gigabit_ethernet	highSpeedFormOf	Ethernet
Iron_python	adaptationOf	Python
Barack_Obama	supports	Military_Action_in_Libya
Donald_Trump	opposes	Illegal_immigration
UK	supports	Brexit
Scotland	opposes	Brexit
Pseudonomas	foundIn	Sea
Pseudonomas	actsOn	Acetone
Hydrocarbon_degradation	occurIn	Sea
Hydrocarbon_degradation	carriedOn	Acetone

Determining the nature of microbes

- **Given**: corpus of research articles
- **Goal**: determine all microbes which are aerobic/anaerobic

- **Challenge**: Cannot extract "isAerobic" directly
- **Method**: Extract triples from the corpus and organise the information, *reason* about the "isAerobic" releation

Pseudonomas	foundIn	Sea
Pseudonomas	actsOn	Acetone
Hydrocarbon_degradation	occurIn	Sea
Hydrocarbon_degradation	carriedOn	Acetone

Interesting entity types
Microbes
Substrate
Enzyme
Process
Environment
Nutrient
Property

Outline

- Introduction
- Applications
- Construction
- Conclusion

General principles for domain-specific KB construction

Source identification

- Structured sources
  - Wikipedia, IMDB, domain-specific websites such as Webopedia…
- Unstructured sources
  - Newspaper articles, blogs, forums, reviews, tweets, comments…

Entities

- Extraction from structured sources
  - Design of appropriate heuristics and/or wrappers
- Extraction from unstructured sources
  - POS tagging + heuristics
- Types known/unknown
- Cleanup
  - Entity resolution to reconcile duplicate entity extractions

Relations

- From both structured and unstructured sources:
  - Known relations
  - Unknown relations
- Cleanup
  - Identifying and reconciling duplicate relationship names

Refinement

- "Complete" the KB by inferring new relations.
  - Simple string matching sometimes does the trick!
  - Graph embeddings are the method of choice (hype)
- Enhance the KB through feedback

Using the general principles

	Sources	Entity extraction techniques	Relation extraction techniques	Refinement
Opinion-base [WSDM 2012, CIKM 2012, CIKM workshops 2012]	Newspaper articles, Wikipedia, Debatepedia	1. Names of entities from YAGO, Debatepedia, *known types* 2. Phrase extraction for topics, *unknown types*	Known relations ("support"/"oppose"), unstructured sources	KB enhancement by feeding back new surface patterns for the relations
TeKnowbase [WWW workshops 2018, ISWC 2020, ECIR 2020, ECIR 2021]	Wikipedia, Webopedia, Techtarget, online textbooks	1. Wikipedia, Webopedia, TechTarget article names, *unknown types* 2. Indexes of online textbooks, *unknown types*	1. Known relations, structured sources 2. KR, unstructured sources 3. Unknown relations, SS 4. UR, US	1. Sub-string extraction • "Graph algorithm" 2. Embeddings for knowledge-base completion
Biochemical KB [Report available on request]	Research articles from the domain, BRENDA, MicrobesOnline, etc.	1. Expert-curated list 2. Web-sources, *known types* 3. POS-tagging + heuristics from the research articles	1. Known relations, unstructured sources • Training ML models after acquiring examples	None
Event-base	Newspaper articles	1. Unknown entities, therefore, POS-tagging + heuristics	1. Requires *nested triples*. 2. Partially known relations, unstructured sources	None End

OpinioNet

Dictionary creation

Relation extraction

Entity and facet extraction

Subtopics

Hamid Karzai

Article Talk

From Wikipedia, the free encyclopedia

"Karzai" redirects here. For the surname, see Karzai (surname).

Hamid Karzai (/ˈhæmɪd ˈkɑːrzaɪ/; Pashto/Persian: حامد کرزی, Pashto pro
pronunciation: [ˈhɔːmɪd karzeˈiː]; born 24 December 1957) is an Afghan s

Afghan peace process

Article Talk

From Wikipedia,

Peace process

1978 Saur Rev

Disarmament, demobilisation and reintegration (DDR), or **disarmament demobilisation, repatriation, reinteg and resettlement (DDRRR)** are strate used as a component of peace proces

Stake-holders, Topics

Afghan President Says **He Supports** Talks With Taliban : The Two ...

www.npr.org/.../afghan-president-says-**he-supports**-talks-with-taliban

4 Jan 2012 – Experts say the plans for talks are a positive step toward a future peace in Afghanistan.

Known relations, unstructured source

Hamid_Karzai

Afghan President

Karzai

Source	Overall	Support	Oppose	ALJ	BBC	CNN	GN	NY	WP
#Opinions	29648	16011	13637	970	3349	2650	6861	9877	5941
#Evaluated	2005	1121	884	364	349	308	301	364	319
Precision	0.724	0.70	0.75	0.72	0.66	0.78	0.788	0.74	0.69

Table 3.8: Opinions & precision results

War or

Afghanistan_War

Iraq_War

Afghan_Peace_Process

Drone_ Strikes

Suicide_Bombings

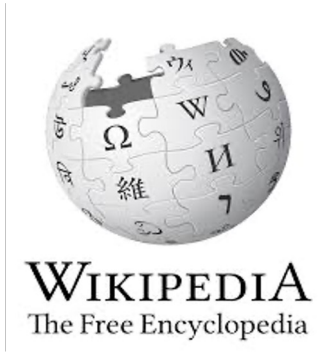
End

Constructing TeKnowbase: Dictionary of entities

WEBOPEDIA™ Everything you need to know is right here.

TechTarget®

24,000



54,000

16,500

Online textbooks

Category: Computing

70,000

TeKnowbase [WWW Workshops 2019]

*Dataset available from https://github.com/prajnaupadhyay/TeKnowbase

Structured sources, known relations

typeOf

List of data structures

Trees [edit]

Main article: Tree (data structure)

Binary trees [edit]

- AA tree
- AVL tree
- Binary search tree
- Binary tree
- Cartesian tree
- Left-child right-sibling binary tree
- Order statistic tree
- Pagoda
- Randomized binary search tree
- Red-black tree

Binary_search_tree	typeOf	Binary_tree
Binary_tree	typeOf	Tree
Tree	typeOf	Data_structure

subtopicOf

List of set theory topics

Articles on individual set theory topics

- Algebra of sets
- Axiom of choice
  - Axiom of countable choice
  - Axiom of dependent choice
  - Zorn's lemma
- Boolean-valued model
- Burali-Forti paradox
- Cantor's back-and-forth method

Boolean-valued_model	subtopicOf	Set_theory
Algebra_of_sets	subtopicOf	Set_theory

applicationOf

Coding theory

From Wikipedia, the free encyclopedia

Contents [hide]

1 History of coding theory
2 Source coding
  2.1 Definition
  2.2 Properties
  2.3 Principle
  2.4 Example
3 Channel coding
  3.1 Linear codes
    3.1.1 Linear block codes
    3.1.2 Convolutional codes
4 Cryptographical coding
5 Line coding
6 Other applications of coding theory
  6.1 Group testing
  6.2 Analog coding

Group_testing	applOf	Coding_theory
Analog_coding	applOf	Coding_theory

algorithmFor

Derivative-free optimization

Algorithms [edit]

- DONE
- Bayesian optimization
- Coordinate descent and adaptive coordinate d
- Cuckoo search
- Evolution strategies, Natural evolution strategi SNES)
- Genetic algorithms
- LIPO algorithm
- MCS algorithm
- Nelder-Mead method

Cuckoo_search	algoFor	Derivative-free_opt
MCS_algorithm	algoFor	Derivative-free_opt

techniqueIn

Document classification

Techniques [edit]

Automatic document classification techniques include:

- Expectation maximization (EM)
- Naive Bayes classifier
- tf–idf
- Instantaneously trained neural networks
- Latent semantic indexing
- Support vector machines (SVM)
- Artificial neural network
- K-nearest neighbour algorithms
- Decision trees such as ID3 or C4.5

Naïve-Bayes_classifier	techIn	Document_classification
K-nearest_neighbor_algorithm	techIn	Document_classification

More relation extractions

Structured Source, Unknown relation

Template:Databases

From Wikipedia, the free encyclopedia

V • T • E

Types	Object-oriented (comparison) • Relational (comparison) • Do
Concepts	Database • ACID • Armstrong's axioms • CAP theorem • CRU
Objects	Relation (table • column • row) • View • Transaction • Transa
Components	Concurrency control • Data dictionary • JDBC • XQJ • ODBC •

View	Object	Databases
Data_dictionary	Component	Databases

Untructured Source, Known relation

synonymOf

"is abbreviation for"

"bfs (breadth-first_search)"

"is short for"

BFS	synonymOf	Breadth-first_search
RegEx	synonymOf	Regular_expression

Untructured Source, Unnown relation

Step	Action	No. of triples	Problem
1	Run OpenIE	400,000	Non-technical entities in the result
2	Remove entities not in dictionary	300,000	Long phrases containing entity
3	Retain triples with 50% match	100,000	Generic entities
4	Remove triples where argument start with "The"	3520	Accuracy still around 65%

gigabig_ethernet	is_a_high_speed_form_of	ethernet
ironpython	is_an_adaption_of	python
utc	uses	gregorian_calender

Refinement using KG embeddings

Xbasic	Typeof	Programming_language
Binomial_heap	Typeof	Tree
Palmdos	Typeof	Operating_system
Delayed_column_generation	Typeof	Convex_programing

TeKnowbase statistics

No. of unique entities	70,285
No. of unique relations	2,574
Taxonomic relations (`typeOf`)	27,078
Total no. of triples	146,657
No. of overlapping entities with DBPedia	17,987
No. of overlapping entities with Freebase	34,785
No. of triples extracted from Wikipedia	99,357
No. of triples extracted from Unstructured sources	3,506

#	**Relation** (rows 1–5)	**# Evaluated triples**	**Accuracy**
1.	`typeOf`	515	99.0% ± 0.8%
2.	`terminologyOf`	676	98.9% ± 0.7%
3.	`synonymOf`	70	100% ± 0.0%
4.	`subTopicOf`	42	91.3% ± 8.2%
5.	`conceptIn`	334	95.4% ± 2.1%
6.	*Unstructured sources*	435	63.2% ± 3.7%
7.	*Inferencing with NTN*	428	64.2% ± 4.5%

Conclusion

- Applications
  - Opinion-base (*heat maps, correlated opinions, flip-floppers*)
  - TeKnowbase (*academic search, pre-requisite construction*)
  - Biochemical KB (*reasoning – incomplete*)
- Construction
  - General principles (*most important: identify good sources, construct a dictionary of interesting entities*)

Thanks: Rawia Awadallah, Srikanta Bedathur, Tanmoy Chakraborty, Prajna Upadhyay, Manas Joshi, Shubham Singla, Ashutosh Bindal, Manjeet Kumar, Gerhard Weikum