Hierarchical Classification of Web Content

Susan Dumais
Microsoft Research

Hao Chen
University of California at Berkeley

SIGIR 2000

Presented by Xudong He, Ailun Yi

University of Edinburgh
March, 2009

Outline

- Background
- Text Categorisation
- Hierarchical Classification
- Data Set and Feature Selection
- SVM
- Result and Evaluation
- Conclusion

Background

- Search Results Representation

Background

- Search Results Representation
- Flat ranked list. e.g. Google

Background

- Search Results Representation
- Flat ranked list. e.g. Google
- Clustering, Clusty
  - Automatic but hard to assign labels
  - Runtime cost, hard for large set of documents

Background

- Search Results Representation
- Flat ranked list. e.g. Google
- Clustering, Clusty
  - Automatic but hard to assign labels
  - Runtime cost, hard for large set of documents
- Hierarchical, Directory-style. ODP, Yahoo! Directory

Background

- Search Results Representation
- Flat ranked list. e.g. Google
- Clustering, Clusty
  - Automatic but hard to assign labels
  - Runtime cost, hard for large set of documents
- Hierarchical, Directory-style. ODP, Yahoo! Directory
- Motivation: Combine both?
  - Need automatic text classifiers
  - Efficiency Consideration – 16+ top level categories, thousands lower-level ones

Text Categorisation

Text Categorisation

- Classify documents into broad semantic topics (e.g. Politics, Entertainment, Sports, etc.)

Democratic vice presidential candidate John Edwards on Sunday accused President Bush and Vice President Dick Cheney of misleading Americans by implying a link between deposed Iraqi President Saddam Hussein and the Sept. 11, 2001 terrorist attacks.

While No. 1 Southern California and No. 2 Oklahoma had no problems holding on to the top two spots with lopsided wins, four teams fell out of the rankings — Kansas State and Missouri from the Big 12 and Clemson from the Atlantic Coast Conference and Oregon from the Pac-10.

Text Categorisation

- Classify documents into broad semantic topics (e.g. Politics, Entertainment, Sports, etc.)

Democratic vice presidential candidate John Edwards on Sunday accused President Bush and Vice President Dick Cheney of misleading Americans by implying a link between deposed Iraqi President Saddam Hussein and the Sept. 11, 2001 terrorist attacks.

While No. 1 Southern California and No. 2 Oklahoma had no problems holding on to the top two spots with lopsided wins, four teams fell out of the rankings — Kansas State and Missouri from the Big 12 and Clemson from the Atlantic Coast Conference and Oregon from the Pac-10.

- Supervised Learning Problem
  - Common algorithms: SVM, NB, kNN, NNet

Text Categorisation Application

Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare

Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare
- News Stories / Academic Paper / Blogs / Tweets, by Subject

Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare
- News Stories / Academic Paper / Blogs / Tweets, by Subject
- Selective Dissemination of Information to information consumers

Text Categorisation Application

- Familiar Spam/Ham, Writing style – female / male, likeliness to Shakespeare
- News Stories / Academic Paper / Blogs / Tweets, by Subject
- Selective Dissemination of Information to information consumers
- Natual Language Processing (Word Sense Disambiguation, POS Tagging etc.)

Hierarchical Classification

Hierarchical Classification

- Multi-class problem
  - 1-againt-rest, sort ***p(c | d)*** into categories

Hierarchical Classification

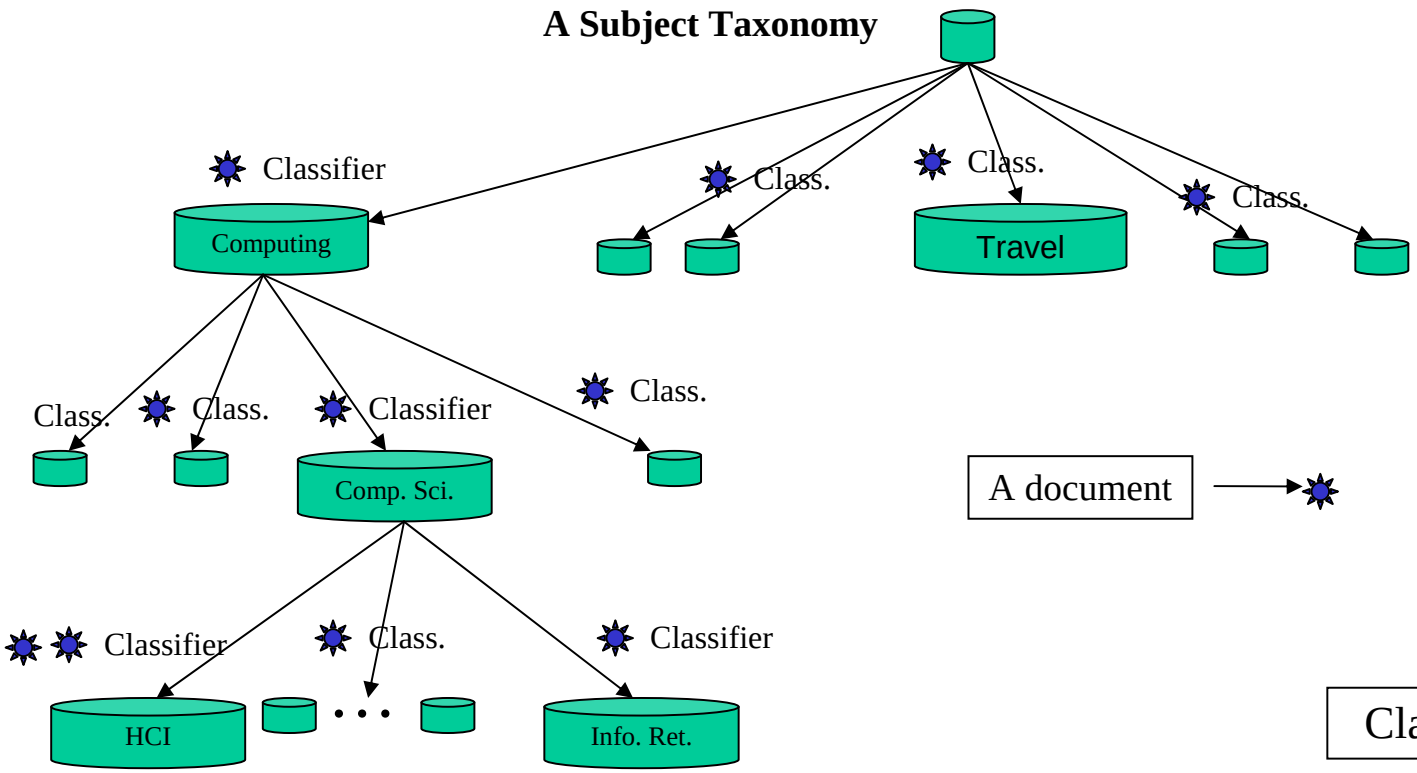
- Multi-class problem
  - 1-againt-rest, sort ***p(c | d)*** into categories
- Multi-level problem
  - How?

Hierarchical Classification

- Multi-class problem
  - 1-againt-rest, sort ***p(c | d)*** into categories
- Multi-level problem
  - How?
- Our Goal
  - Classify into Hierarchical Categories (for Search Results)

The Approach: Flat vs. Hierarchical

The Approach: Flat vs. Hierarchical



A document →

Classifier → Confidence the document belongs to the class

Comparison among
(1) P(L2)
(2) P(L1) && P(L2)
(3) P(L1) * P(L2)

Experiment Setting: Data Set

- LookSmart directory (May, 1999)
  - 370,597 web pages
  - 17,173 categories over 7-level hierarchy
- Focus on top 2 levels
  - 13 top-level & 150 second-level categories
- Train / Test split
  - Randomly select 16%, 50,078 pages for training
  - 10,024 pages for testing

Experiment Setting: Feature Preprocessing

- Feature Vector
  - Words (uni-gram vs. bigram)
  - Binary (vs. Tf-idf)
- Feature Selection
  - Top 1000 by Mutual Information:

$$MI(F,C) = \sum_{F \in \{f,\bar{f}\}} \sum_{C \in \{c,\bar{c}\}} P(F,C) \log \frac{P(F,C)}{P(F)P(C)}$$

- Second-level Local Selection
  - Second level feature sets narrowed by top-level categories, more balanced features

The SVM Model

- The equation of a hyperplane:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}+b=0$$

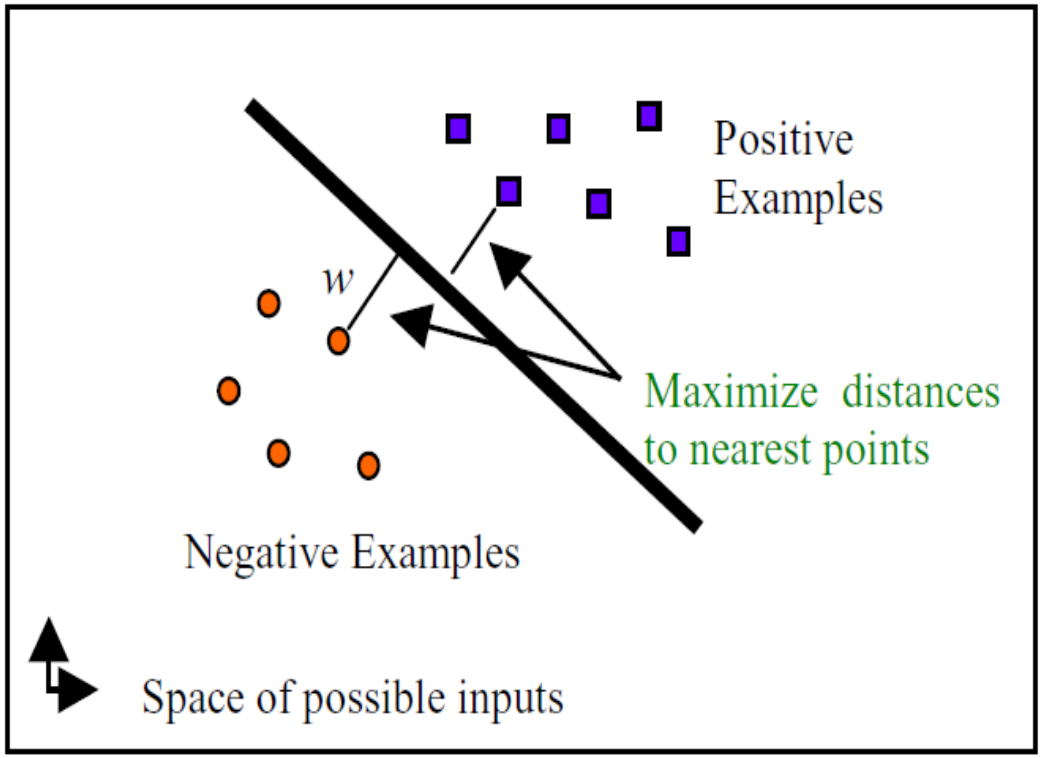
- The separator property:

$$\mathbf{w}^T\mathbf{x_i}+b\geq 1 \text{ iff } y_i=+1$$

$$\mathbf{w}^T\mathbf{x_i}+b\leq 1 \text{ iff } y_i=-1$$

- Minimise $\|\mathbf{w}\|^2$ subject to:

$$\forall_{i=1}^{n}: y_i\left[\mathbf{w}^T\mathbf{x_i}+b\right]\geq 1-\xi_i$$



- The decision function:

$$f(x)=\mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i\cdot y_i\cdot(\mathrm{x_i}\cdot\mathrm{x})+b\right)$$

- Independent of the dimensionality of the feature space

The normal of the decision surface is a linear combination of examples

Why SVMs are good for TC

Why SVMs are good for TC

- High-dimensional input space.

  Use overfitting protection not dependent on the number o features

- Few irrelevant features:

  almost all feature contain considerable information.

Why SVMs are good for TC

- High-dimensional input space.

  Use overfitting protection not dependent on the number o features

- Few irrelevant features:

  almost all feature contain considerable information.

- Sparse document vectors

  Each of the document vectors contain only a few non-zero element

- Most text categorisation problems are linearly separable

Train and Optimise the SVM

Train and Optimise the SVM

- Use the simplest linear SVM

  good classification accuracy

  fast to learn and apply

- Use Platt's sequential minimal optimization(SMO) algorithm

  Break large quadratic programming(QP) problem into small QP problems

Train and Optimise the SVM

- Use the simplest linear SVM

  good classification accuracy

  fast to learn and apply

- Use Platt's sequential minimal optimization(SMO) algorithm

  Break large quadratic programming(QP) problem into small QP problems

- New items are classified by using **w.x**.

- Fit a sigmoid to output by maximum likelihood and produce posterior probabilities.

Train and Optimise the SVM

- Use the simplest linear SVM

  good classification accuracy

  fast to learn and apply

- Use Platt's sequential minimal optimization(SMO) algorithm

  Break large quadratic programming(QP) problem into small QP problems

- New items are classified by using **w.x**.

- Fit a sigmoid to output by maximum likelihood and produce posterior probabilities.

- c, the penalty imposed on training instances.

  Default value 0.01.

- p, the decision threshold to control precision and recall

  Flat non-hierarchical models, p=0.5

  Hierarchical models with multiplicative decision rule, p=0.2

  For the Boolean decision rule, p_1=0.2 (first level) and p_2=0.5 (second level)

Experimental Results and Evaluation [1]

Experimental Results and Evaluation [1]

- Accuracy

  Measure: F1 = 2PR/(P+R) P: precision, R: recall

Experimental Results and Evaluation [1]

- Accuracy

Measure: F1 = 2PR/(P+R)	P: precision, R: recall	
Top level	.572	
By optimizing the C parameter, 15%~20% improvement		
Non-hierarchical	.476	baseline

Experimental Results and Evaluation [1]

- Accuracy

Measure: F1 = 2PR/(P+R)	P: precision, R: recall	
Top level	.572	
By optimizing the C parameter, 15%~20% improvement		
Non-hierarchical	.476	baseline
P(L1)*P(L2)	.495	4% improvement
P(L1)&&P(L2)	.497	4.4% improvement

No difference between Boolean rule and multiplicative scoring rule

Experimental Results and Evaluation [1]

- Accuracy

Measure: F1 = 2PR/(P+R)	P: precision, R: recall	
Top level	.572	
By optimizing the C parameter, 15%~20% improvement)		
Non-hierarchical	.476	baseline
P(L1)*P(L2)	.495	4% improvement
P(L1)&&P(L2)	.497	4.4% improvement
No difference between Boolean rule and multiplicative scoring rule		
Assuming P(L1)=1	.711	49.34% !!!

Best and Worst F1 Score by Category

Best and Worst F1 Score by Category

	Category Name
Best F1	
0.841	*Health & Fitness/Drugs & Medicines*
0.797	*Home & Family/Real Estate*
0.781	*Reference & Education/K-12 Education*
0.750	*Sports & Recreation/Fishing*
0.741	*Reference & Education/Higher & Cont. Ed.*
Worst F1	
0.034	*Society & Politics/World Cultures*
0.088	*Home & Family/For Kids*
0.122	*Computers & Internet/News & Magazines*
0.131	*Computers & Internet/Internet & the Web*
0.133	*Business & Finance/Business Professions*

Best and Worst F1 Score by Category

	Category Name
Best F1	
0.841	*Health & Fitness/Drugs & Medicines*
0.797	*Home & Family/Real Estate*
0.781	*Reference & Education/K-12 Education*
0.750	*Sports & Recreation/Fishing*
0.741	*Reference & Education/Higher & Cont. Ed.*
Worst F1	
0.034	*Society & Politics/World Cultures*
0.088	*Home & Family/For Kids*
0.122	*Computers & Internet/News & Magazines*
0.131	*Computers & Internet/Internet & the Web*
0.133	*Business & Finance/Business Professions*

- Performance varies widely across categories
- Difficult categories are based on non-content distinctions.

Experimental Results and Evaluation [2]

- Efficiency
  - Offline training

    Measured by the CPU seconds of the standard 266MHz Pentium Ⅱ PC running WinNT

    The total training time for all the models in the paper was only about 30 minutes

For all the non-hierarchical models on 50078 examples	729
For all 150 hierarchical second-level models	128
For the 13 top-level categories	1258

Experimental Results and Evaluation [2]

- Online evaluation

  focus on the cost of operation on the dot product of eight vector and the feature vector.

Models	Each instance Compared to
Non-hierarchical	150 second-level categories
Hierarchical, the multiplicative scoring rule	13 first-level and all 150 second-level categories
Hierarchical, the Boolean scoring rule	all 13 first-level **but only** the second-level categories pass the first-level criterion

More speed gain in the Boolean decision rule, overall 14.8% of the multiplicative case, 16% of the non-hierarchical case.

Yet another Hierarchical SVM (vs. Flat)

Yet another Hierarchical SVM (vs. Flat)

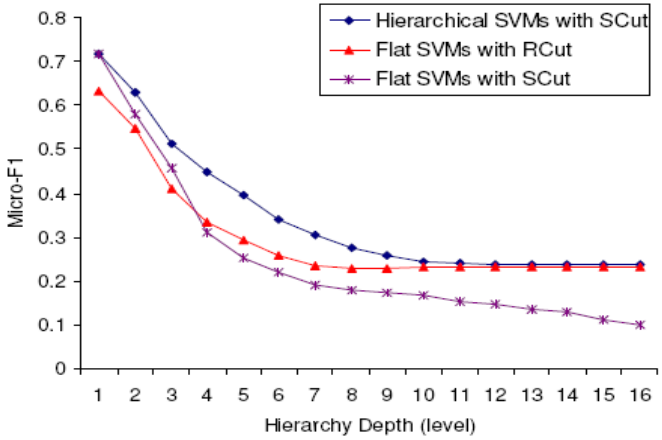
- Y Yang et al (2005) investigated the efficiency and effectiveness of SVMs over very large-scale taxonomies:
  - Yahoo! Directory: classify 792,601 documents into 292,216 categories within 16 levels;
  - Cluster of SVMs

    10 x 3GHz CPU and 4GB Ram

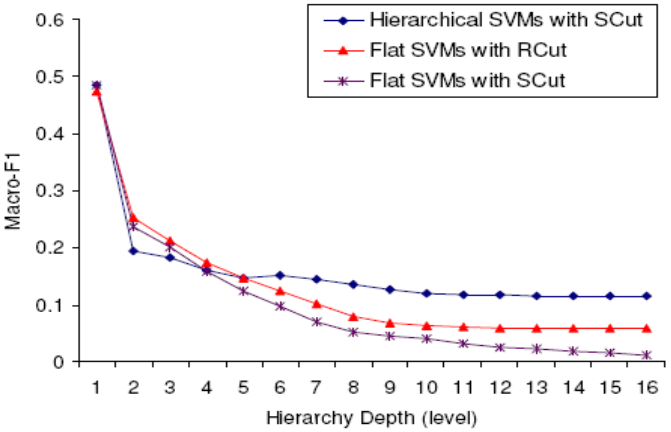
Yet another Hierarchical SVM (vs. Flat)

- Y Yang et al (2005) investigated the efficiency and effectiveness of SVMs over very large-scale taxonomies:
  - Yahoo! Directory: classify 792,601 documents into 292,216 categories within 16 levels;
  - Cluster of SVMs

    10 x 3GHz CPU and 4GB Ram
- Effectiveness
  - Benefit balanced local feature sets



(a)



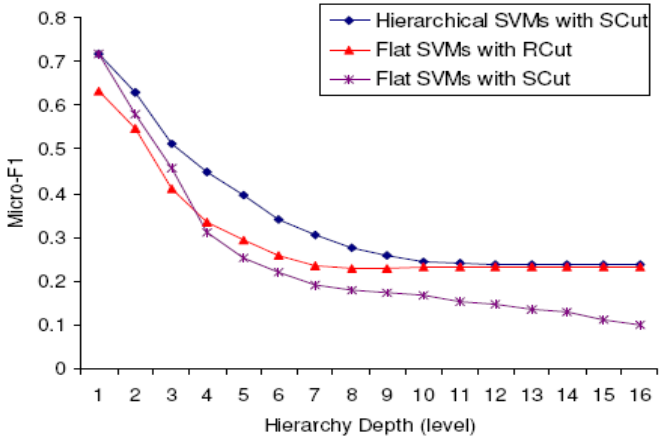
(b)

Yet another Hierarchical SVM (vs. Flat)

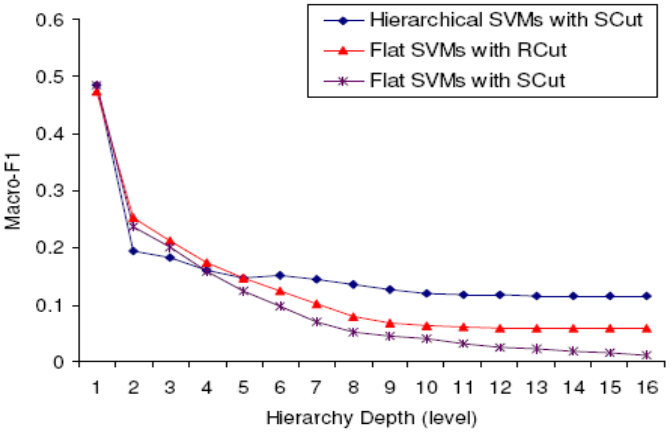
- Y Yang et al (2005) investigated the efficiency and effectiveness of SVMs over very large-scale taxonomies:
  - Yahoo! Directory: classify 792,601 documents into 292,216 categories within 16 levels;
  - Cluster of SVMs

    10 x 3GHz CPU and 4GB Ram

- Effectiveness
  - Benefit balanced local feature sets
- Efficiency
  - Save 90% training time



(a)



(b)

Time complexity of SVMs (h)

Algorithms	Training		Testing
	without SCut	with SCut	with SCut
Hierarchical SVMs	0.42	2.10	0.12
Flat SVMs	310.87	1304.64	53.63

Conclusion

Conclusion

- Novel approach, much more efficient on large taxomony
- Balance feature sets, better effectiveness
- Future works: generalised sequential model in hierarchical classification

Conclusion

- Novel approach, much more efficient on large taxomony
- Balance feature sets, better effectiveness
- Future works: generalised sequential model in hierarchical classification
- Questions?

Reference

- Susan Dumais, Hao Chen, Hierarchical Classification of Web Content, Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, 2000.
- Yiming Yang , Xin Liu, A re-examination of text categorization methods, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.42-49, August 15-19, 1999, Berkeley, California, United States.
- Y Yang et al, Support vector machines classification with a very large-scale taxonomy, ACM SIGKDD Explorations Newsletter, 2005
- T Joachims, Text categorization with support vector machines: learning with many relevant features, ECML-98, 1998

- Thank you

web news images wikipedia blogs jobs more »

Clusty

Jaguar | Search | advanced preferences

clusters sources sites

remix

All Results (233)

- Jaguar Cars (36)
- Parts (44)
- Photos (34)
- Club (36)
- Panthera onca (18)
- Animal (14)
- Leopard (12)
- Jacksonville Jaguars (9)
- Autos, Research (6)
- 2010 Jaguar (5)

more | all clusters

find in clusters: Find

Font size: A A A A

Top **228** results of at least **2,561,000** retrieved for the query **Jaguar** (definition) (details)

Sponsored Results

Jaguar ™ - Official Site - Home Of The Award Winning **Jaguar** XF Voted Best Executive Car Of 2009 - www.Jaguar.co.uk

Approved **Jaguar** Dealer - New & Approved Used **Jaguar** Sales With National Delivery. Reading. - www.Reading-Jaguar.co.uk

Car Finder Scotland - More Scottish used cars than any other website! - www.carfinderscotland.co.uk

Search Results

1. **Jaguar**
   Official worldwide web site of **Jaguar** Cars. Directs users to pages tailored to country-specific markets and model-specific websites.
   www.jaguar.com - [cache] - Live, Open Directory, Ask

2. **Jaguar** US - Home
   Jaguar USA official website
   www.jaguarusa.com - [cache] - Live, Ask

3. **Jaguar** - Wikipedia, the free encyclopedia
   The **jaguar**, Panthera onca, is a New World feline and one of four "big cats" in the Panthera genus, along with the tiger, lion, and leopard of the Old World. It is the only Panthera found in the New World. The **jaguar** is the third-largest feline after the tiger and the lion, and on average the largest and most powerful feline in the Western Hemisphere.Etymology · Taxonomy · Biology and behavior · Ecology · Conservation status
   en.wikipedia.org/wiki/Jaguar - [cache] - Live, Ask

4. **Jaguar** - Defenders of Wildlife
   Get the facts about **jaguars**. ... **Jaguar**. Panthera onca. The **jaguar** is the largest cat in the Americas. According to one Indian myth, the **jaguar** acquired its spotted coat by daubing ...
   www.defenders.org/wildlife_and_habitat/wildlife/jaguar.php - [cache] - Live, Ask

5. Jacksonville **Jaguars**
   The official team site with scores, news items, game schedule, and roster.
   www.jaguars.com - [cache] - Open Directory, Ask

6. Rare Arizona **Jaguar** Euthanized
   Mar 4, 2009 - A rare **jaguar** captured and collared in Arizona two weeks ago was euthanized after falling ill, state game officials said. The **jaguar** was the only one spotted in the United States in more than a decade. Officials captured the cat Feb. 18 as part of ... mountain lions. A procedure had previously been put in place to attach a satellite collar to any such fortuitously captured **jaguar**, so wildlife experts could monitor its movements to learn more about the extremely rare creatures. The **jaguar** stopped moving, however. ...
   news.yahoo.com/s/livescience/rarearizonajaguareuthanized - [cache] - Yahoo! News

7. **Jaguar AU - Jaguar** Cars
   Information on new, preowned, services and news on models.
   www.jaguar.com.au - [cache] - Live, Ask, Open Directory

dmoz open directory project

In partnership with AOL search

about dmoz | dmoz blog | suggest URL | help | link | editor login

Search advanced

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web

Copyright ©1998-2009 Netscape

4,570,989 sites - 82,392 editors - over 590,000 categories