Classification of Web Search Results

A talk by

Nishant Mittal
Graduate Student, Computer Science and Engineering, University of California, San Diego
n mittal@cs.ucsd.edu
References

- **Hierarchical Classification of Web Content (2000)**
  - Susan Dumais, Microsoft Research
  - Hao Chen, Computer Science Division, University of California Berkeley

- **Optimizing Search by Showing Results In Context (2001)**
  - Susan Dumais and Edward Cutrell, Microsoft Research
  - Hao Chen, Computer Science Division, University of California Berkeley

- **Text Classification for Web Filtering (Lecture Notes, 2004)**
  - Fabrizio Sebastiani
Overview of this talk

- Introduction
- How is hierarchical classification different from non-hierarchical classification?
- Classification approach
- Results
- Illustrate advantages of hierarchical classification
Introduction

- As the Internet continues to grow, information retrieval is becoming increasingly difficult.

- Search engines (Google, AltaVista…) and web directories (Yahoo! …) address this problem but are far from a solution.
- Results shown based purely on rank.

- Search results are without any context.

- Intermixed results.

- Chunk of irrelevant results before the desired one.
Even for a seemingly specific search, the number of results confuses the user.

There is no “Live Concert” related link in the top 25 results.
Existing Solutions

Clustered Results

1. Floydian Slapton: The Pink Floyd Experience > Albums > David Gilmour > David Gilmour
   All about the album. Photos, song list, sound clips and links to purchase it. Floyd guitarist David Gilmour embarked on his first solo effort following... reunion group Gilmour was in before joining Pink...
   Source: floydianslapton.com/dis/davidgil.htm - show in cluster

2. DAVE GILMOUR (new window) [frame] [preview]
   DAVE GILMOUR, Gilmour and Pink Floyd. Dave Gilmour is one of the world's most popular guitarists. As part of Pink Floyd and...
   Source: funkereurope.com/players/DAVEGILMOUR.asp - show in cluster

3. Pink Floyd - Gen's Page (new window) [frame] [preview]
   Fan has assembled links to media from all over the Web, including sound and video files in multiple formats.
   Source: gencities.com/SunsetStripVenue/6162/index.html - show in cluster

4. DAVE GILMOUR WEBCAST at All Pink Floyd Fan Network! - Pink Floyd... (new window) [frame] [preview]
   DAVE GILMOUR WEBCAST - All Pink Floyd Fan Network - The best place on Pink Floyd and Roger Waters, David Gilmour, Syd Barrett, Rick Wright and Nick Mason.
   Source: pinkfloydfan.net/817-15-1.htm - show in cluster

5. AC/DC Chris Slade Solo - Dave Gilmour of Pink Floyd's About Face... (new window) [frame] [preview]
   AC/DC Chris Slade Solo - Dave Gilmour of Pink Floyd's About Face Tour 1994. AC/DC Chris Slade Solo - Dave Gilmour of Pink Floyd's...
   Source: cradbodyinblue.com/acdcsaboutface.htm - show in cluster

6. David Gilmour (new window) [frame] [preview]
   David Gilmour joined Pink Floyd in 1968 in order that the band might have a conscious guitarist for... to write and record songs, and get Dave to play the show...
   Source: musi17, lyrics18

7. Dave Gilmour / Michael Kamen / AWSO P? (new window) [frame] [preview]
Classification Vs Clustering

- Clustering generates categories “on the fly,” then classifies search results into these categories.
- Clustering eliminates the tedious task of developing and maintaining a list of categories.
- Computationally intensive.

But…

- Classification is the process of assigning documents to pre-defined categories.
- Classification of documents based on already learned features is very efficient.
- Manually generated category names are easily understood.
Classification - Definition

- Assign natural language documents to pre-defined categories.

- Membership of a document in a category cannot be decided with certainty.
Assumptions

- Categories are just symbolic labels. No additional knowledge of their meaning is available to help in building the classifier; in particular, the text of the label is insignificant.

- The attribution of documents to categories should be done on the basis of the content of the document and not on the basis of prior knowledge about the document.
Types of Classification

**HARD** classification
- Come up with a value that indicates membership or non-membership of $d_j$ in $c_i$

**SOFT** classification
- Come up with a value in [0,1] which indicates the degree of confidence of the system in the membership of $d_j$ in $c_i$
Hierarchical Vs Non-Hierarchical

- In the hierarchical case, a model is learned to distinguish a second-level category from other categories within the same top level.

- In the flat non-hierarchical case, a model distinguishes a second-level category from all other second-level categories.
Importance of Hierarchy

- Many potentially good features are not useful discriminators in non-hierarchical representations (Computers)

- Helps at second-level (Chat)

- Does not handle errors in related categories (Soccer and Football)
Classification Approach

- **Training/Test Selection:** Train classifiers for each category based on a set of pre-categorized documents.

- **Pre-processing:** Remove noise from the document.

- **Feature Selection:** Come up with a list of features (words) that really define the document.

- **SVM Parameters:** Define decision thresholds for various levels in the hierarchy.
Classifiers

- Each category has its own classifier.

- A document is either in the category or not.

- Support Vector Machines are ideally suited for this One-Vs-All Scenario.
Training/Test Selection

LookSmart Web Directory

- Large heterogeneous collection of pages
- Consists of 370597 unique pages, manually classified into a hierarchy of categories
- A 7-level hierarchy with 17,173 categories
- Only the top two categories were considered

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Total</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automotive</td>
<td>4982</td>
<td>578</td>
<td>114</td>
</tr>
<tr>
<td>Business &amp; Finance</td>
<td>31599</td>
<td>3508</td>
<td>703</td>
</tr>
<tr>
<td>Computers &amp; Internet</td>
<td>46000</td>
<td>5718</td>
<td>1126</td>
</tr>
<tr>
<td>Entertainment &amp; Media</td>
<td>88697</td>
<td>11163</td>
<td>2159</td>
</tr>
<tr>
<td>Health &amp; Fitness</td>
<td>25380</td>
<td>3500</td>
<td>722</td>
</tr>
<tr>
<td>Hobbies &amp; Interests</td>
<td>22959</td>
<td>3227</td>
<td>682</td>
</tr>
<tr>
<td>Home &amp; Family</td>
<td>11484</td>
<td>1373</td>
<td>280</td>
</tr>
<tr>
<td>People &amp; Chat</td>
<td>35157</td>
<td>3309</td>
<td>682</td>
</tr>
<tr>
<td>Reference &amp; Education</td>
<td>58002</td>
<td>5574</td>
<td>1175</td>
</tr>
<tr>
<td>Shopping &amp; Services</td>
<td>19667</td>
<td>2122</td>
<td>423</td>
</tr>
<tr>
<td>Society &amp; Politics</td>
<td>38968</td>
<td>4855</td>
<td>946</td>
</tr>
<tr>
<td>Sports &amp; Recreation</td>
<td>23559</td>
<td>3081</td>
<td>640</td>
</tr>
<tr>
<td>Travel &amp; Vacations</td>
<td>43685</td>
<td>5409</td>
<td>1091</td>
</tr>
<tr>
<td>Total Unique</td>
<td>370597</td>
<td>50078</td>
<td>10024</td>
</tr>
</tbody>
</table>
Pre-processing

- For all web pages the plain text was extracted. The title, META keywords and ALT field from the IMG tags were included (Links were not included).

- A small stop list was used to omit the most common words.

- A binary vector was created for the document indicating whether a term appeared in the page or not (term frequencies were not taken into account).
Feature Selection

- Words appearing in only one document were eliminated.

- Only the top 1000 words with highest mutual information with each category were considered as features.

$$MI(F, C) = \sum_{F \in \{f, \bar{f}\}} \sum_{C \in \{c, \bar{c}\}} P(F, C) \log \frac{P(F, C)}{P(F)P(C)}$$
A linear SVM can be thought of as a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin.

Mathematically...

\[
\text{minimize } \frac{1}{2} \|w\|^2 \text{ subject to } y_i (\bar{w} \cdot \bar{x}_i - b) \geq 1, \forall i
\]
Why SVMs?

- Very efficient, once trained.
- Found to be more accurate for text classification than naïve Bayes and Neural Nets.
- Can be modified to work for non-linearly separable group of documents.
Computing Probability

- After training, a sigmoid is fitted to the output of the SVM [Platt’s Method]
  - Platt, J., Fast training of Support Vector Machines using sequential minimal optimization.

- This is done to get posterior probabilities of a document being in the category under consideration.
Classification Process

- For each document, the probability of it being in each of the top level and second level categories is computed.

- If the probability exceeds the decision threshold, the document is judged to be in the category.

- A document could be in zero, one or many categories.
Decision Threshold

- Defines a lower limit on the probability of a document falling into a category.
- This parameter is different for different levels in the hierarchy.
- The threshold values can be obtained by maximizing performance on the $F_1$ measure.
Measuring Performance

\[ F_1 = \frac{2\pi \rho}{\pi + \rho} \]

\( \pi, \rho \) are micro averaged precision and recall respectively.

**Precision**: Among the classified documents, how many actually belong to the class.

**Recall**: Among all the documents belonging to a class, how many were classified as such.
Combining Probabilities

**Boolean**

\[ P(L1) \&\& P(L2) \]

- We match probabilities on the second level only if the threshold at the top level was matched.
- So the threshold has to be matched at both levels.

**Multiplicative**

\[ P(L1) \times P(L2) \]

- Allows matches even if the scores at one level fall below threshold.
- Computationally demanding as all categories need to be tested.
Results

- Top Level
  - overall performance (F₁) : 0.572

- Second Level (non-hierarchical)
  - overall performance (F₁) : 0.476
  - performance goes down when we increase the number of categories.

- Second Level (hierarchical)
  - Multiplicative (F₁) : 0.495
  - Boolean (F₁) : 0.497
    \[4\% \text{ improvement}\]
Conclusion

- Small improvement in $F_1$ accuracy score by using hierarchical models, compared with the flat non-hierarchical model.

- No difference was found between a multiplicative scoring function and a sequential Boolean function.

- Boolean function is much more efficient as it requires only 14%-16% of the comparisons.
Optimizing Search by showing results in Context

User Study
Query: jaguar
Retrieved 100 documents

- (99) jaguar watches - jaguar watches, tag heuer, omega, rolex, cartier, breitling,
- (98) Jaguar Club of Florida
- (97) Jaguar Underground Dox
- (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B
- (95) Jaguar Clubs of North America
- (94) Welcome to Jaguar
- (93) Terry's Jaguar Parts
- (92) The Jaguar Photo Gallery
- (91) Atari - Jaguar Order Form
- (90) Jaguar Jottings
- (89) Jaguar 64 JUST CLAWS
- (88) Clan Jaguar Quake & Quake 2 Clan
- (87) Jag-lovers Jaguar Cars Windows Wallpaper page
- (86) Jaguar Publishing
- (85) A.C. Imports Jaguar - Palatine IL
- (84) Classic Car Source -- Welsh Jaguar Classic Car Museum
- (83) Atari Jaguar-System
- (82) Jaguar Owners Web Ring
- (81) Jaguar, new cars, used cars, and accessories
- (80) Jaguar Sovereign
- (79) JFPN's Jaguar Chat Intro Page
- (78) The Jaguar Server Home Page
- (77) Westlake Motors Jaguar BMW automobile dealer of Elmhurst, Illinois. New and us
- (76) Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob
List Interface - 2

Query: jaguar
Retrieved 100 documents

- (99) jaguar watches - jaguar watches, tag heuer, omega, rolex, cartier, breitling
  jaguar watches - jaguar watches, tag heuer, omega, rolex, cartier, breitling, seiko, movado, bulova and other fine watch
  http://www.ashford.com/watches/jaguar.htm

- (98) Jaguar Club of Florida
  Welcome to the Jaguar Club of Florida ( JCOF ). This is the official site of the Jaguar automobile club located in central
  http://www.magicnet.net/~jaguar/jcof.html

- (97) Jaguar Underground Doc
  The Jaguar Underground Documentation. (c) Copyright 1996 by Klaus and Nat! Share and spread: YES. Publish and/or sell: NO. This is an attempt to
document.
  http://www.geocities.com/SoHo/l460/dox.html

- (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph
  Renault Peugeot Ferrari Fiat B
  British & European AUTO PARTS: CLASSIC & MODERN, import/export, H.D. Rogers & Sons 3418 Barksdale Blvd, Bossier City Louisiana 71112 USA (318) 742-3651,
  http://www.hdroses.com/

- (95) Jaguar Clubs of North America
  The Jaguar Clubs of North America (JCNA) is the parent organization of 49
corporate Jaguar clubs and serves over 5,000 members in America, Canada and
  http://www.jcna.com/

- (94) Welcome to Jaguar
  Welcome To Jaguar, We Have Models, Scale Figures, Aircraft Accessories,
  Tanks, Resin Kits, War Model Accessories, &
  http://www.jaguarmodels.com/

- (93) Terry's Jaguar Parts
  Terry's Jaguar Home Page. Your one stop for all of your Jaguar XKE, XJ6 and
  XJS needs!
List Interface - 3

Query: jaguar
Retrieved 100 documents

- (99) Jaguar watches - jaguar watches, tag heuer, omega, rolex, cartier, breitling.
  Category: Shopping & Services
  jaguar watches - jaguar watches, tag heuer, omega, rolex, cartier, breitling, seiko, movado, bulova and other fine watch
  http://www.ashford.com/watches/jaguar.htm

- (98) Jaguar Club of Florida
  Category: Automotive
  Welcome to the Jaguar Club of Florida (JCOF). This is the official site of the Jaguar automobile club located in central
  http://www.magicnet.net/~jaguar/jcof.html

- (97) Jaguar Underground Dox
  Category: Not Categorized
  The Jaguar Underground Documentation. (c) Copyright 1996 by Klaus and Nati Share and spread: YES. Publish and/or sell: NO. This is an attempt to
document.
  http://www.geocities.com/SoHo/1460/dox.html

- (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B
  Category: Automotive
  British & European AUTO PARTS- CLASSIC & MODERN, import/export, H.D. Rogers & Sons 3418 Barksdale Blvd, Bossier City Louisiana 71112 USA(318) 742-3651,
  http://www.hdros.com/

- (95) Jaguar Clubs of North America
  Category: Not Categorized
  The Jaguar Clubs of North America (JCNA) is the parent organization of 49 local Jaguar clubs and serves over 5,000 members in America, Canada and
  http://www.jcna.com/

- (94) Welcome to Jaguar
  Category: Shopping & Services
Query: jaguar
Retrieved 100 documents

- Automotive
  - (98) Jaguar Club of Florida
  - (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B
  - (77) Westlake Motors Jaguar BMW automobile dealer of Elmhurst, Illinois. New and us
  - (76) Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob
  - (75) A&L Luxury Car Center - Jaguar Main Page

- Computers & Internet
  - (91) Atari - Jaguar Order Form
  - (88) Clan Jaguar Quake & Quake 2 Clan
  - (87) Jag-lovers Jaguar Cars Windows Wallpaper page
  - (83) Atari Jaguar-System

- Entertainment & Media
  - (92) The Jaguar Photo Gallery
  - (74) Tom's Collection of Jaguar Mark II Photos

- Travel & Vacations
  - (84) Classic Car Source -- Welsh Jaguar Classic Car Museum

- Shopping & Services
Query: jaguar
Retrieved 100 documents

- Automotive
  - (98) Jaguar Club of Florida
    Welcome to the Jaguar Club of Florida (JCOF). This is the official site of the Jaguar automobile club located in central
    http://www.magicnet.net/~jaguar/jcof.html
  - (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph
    Renault Peugeot Ferrari Fiat B
    British & European AUTO PARTS: CLASSIC & MODERN, import/export, H.D.
    Rogers & Sons 3418 Barksdale Blvd, Bossier City Louisiana 71112 USA(318)
    742-3651,
    http://www.hd Rogers.com/
  - (77) Westlake Motors Jaguar BMW automobile dealer of
    Elmhurst, Illinois. New and us
    Westlake Motors Jaguar BMW automobile dealership in Elmhurst, Illinois, near
    Chicago. Check out our new and used car inventory with photos,
    http://www.westlamotors.com/
  - (76) Bauer Jaguar, your specialist in luxury foreign sports cars
    and Jaguar automob
    Southern Californias leading Jaguar dealership for new and select-edition,
    previously-owned automobiles. Full-service capabilities with
    http://www.bauerjaguar.com/
  - (75) A & L Luxury Car Center - Jaguar Main Page
    The largest pre-owned Jaguar and BMW dealership in the Tri-State Area
    http://www.a & l jaguar.com/

- Computers & Internet
  - (91) Atari - Jaguar Order Form
    WE buy manufacturers closeouts so YOU get great savings. You may e-mail
    the order or print it and fax it at (816) 531-6569. Company. Address. City.
    http://www.oshealtd.com/atariorderform.htm
  - (92) Clan Jaguar Onaka & Onaka 2 Clan

Query: jaguar
Retrieved 100 documents

- (98) Jaguar Club of Florida
- (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B
- (77) Westlake Motors Jaguar BMW automobile dealer of Elmhurst, Illinois. New and us
- (76) Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob
- (75) A&L Luxury Car Center - Jaguar Main Page

- (91) Atari - Jaguar Order Form
- (88) Clan Jaguar Quake & Quake 2 Clan
- (87) Jag-lovers Jaguar Cars Windows Wallpaper page
- (83) Atari Jaguar-System

- (92) The Jaguar Photo Gallery
- (74) Tom's Collection of Jaguar Mark II Photos

- (84) Classic Car Source -- Welsh Jaguar Classic Car Museum
Category Interface – 4

Query: jaguar
Retrieved 100 documents

- Automotive SubCateg Show (24)
- Computers & Internet SubCateg Show (21)
- Entertainment & Media SubCateg Show (11)
- Travel & Vacations SubCateg Show (5)
- Shopping & Services SubCateg Show (4)
- Business & Finance SubCateg Show (3)
- Sports & Recreation SubCateg Show (2)
- Hobbies & Interests SubCateg Show (1)
- Society & Politics SubCateg Show (1)
- Not Categorized Show (35)
Results

![Graph showing mean log time for different conditions]

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean Log Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat Hover</td>
<td>73.6</td>
</tr>
<tr>
<td>Cat Inline</td>
<td>70.3</td>
</tr>
<tr>
<td>Cat w/o Names</td>
<td>84.3</td>
</tr>
<tr>
<td>Cat Browse</td>
<td>82.3</td>
</tr>
<tr>
<td>List Hover</td>
<td>104.5</td>
</tr>
<tr>
<td>List Inline</td>
<td>95.8</td>
</tr>
<tr>
<td>List w/ CatNames</td>
<td>98.8</td>
</tr>
</tbody>
</table>
Results

![Bar chart showing mean log time for Easy and Difficult categories and lists.](chart.png)
Conclusion

- Category Interface was found to be better than List Interface.

- For difficult searches, Category Interface was much better than the List Interface.

- Category Interface performed better even when significant context information was removed.

- Inline summaries improved performance.
Classification of Web Content

A talk by

Nishant Mittal
Graduate Student, Computer Science and Engineering,
University of California, San Diego
nmittal@cs.ucsd.edu