

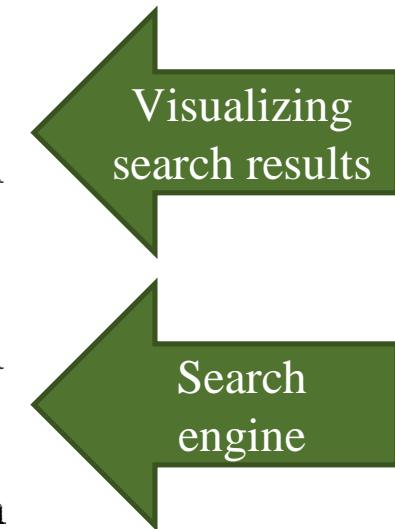
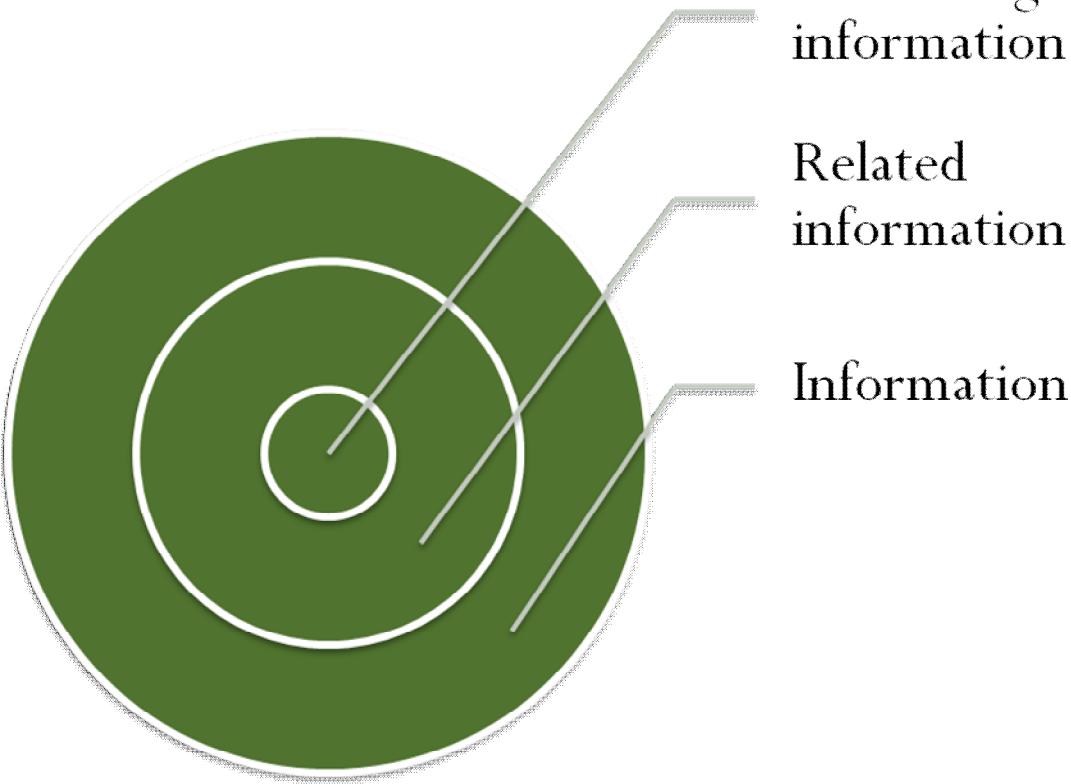
# A Naïve Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Results

Duoqian Miao  
Tongji University  
2011-2-28

# Outline

- Why? (Do search results need to be visualized?)
- How? (How could we realize it?)
- Character of our method
- Bayesian multi-label classification algorithm
- A prototype system: TJ-MLWC
- Conclusion

# Why?



# A example: Google News

Search “China” in Google News

396,695

results!  
Too many!

The screenshot shows the Google News interface. At the top, there is a search bar with 'china' typed into it, a 'Search' button, and links for 'Advanced news search' and 'Preferences'. Below the search bar, the text 'News results: Standard Version | Text Version | Image Version Results 1 - 10 of about 396,695 for china. (0.17 seconds)' is displayed. To the left, a sidebar offers options like 'Recent', 'Last hour', 'Last day', 'Past week', 'Past month', and 'Archives', with 'All dates' at the bottom. The main content area is titled 'Sorted by relevance' with options to 'Sort by date' and 'Sort by date with duplicates included'. It features a news item from ABC News with a thumbnail of the Chinese flag and the headline 'China says US navy ship was breaking law: HK website'. Below this, there are several other news snippets from various sources like Reuters, BEIJING (Reuters), The Associated Press, AFP, Aljazeera.net, and The International Herald Tribune.

There is no  
category with  
related to query  
word!

Displaying all  
search results and  
having no class label.

# A example: Google News (Cont.)

The screenshot shows a Windows Internet Explorer window displaying the Google News homepage. The address bar shows the URL: <http://news.google.com/news/section?cf=all&ned=us&topic=b&ictln>. The left sidebar has a red box around the 'Business' category link. A purple arrow points from a text box containing the explanatory text to this red box. The main content area displays news items under the 'Business' category, with a large text box overlaid containing the explanatory text.

When we don't input query word, it shows news according to category .

In fact, Google only stores news according to a fixed taxonomy. It is not a real-time classification system based on query word given by users.

- Economy
- Society
- Education
- Technology
- Politics
- Military
- Sports
- Entertainment
- Business

Microsoft

Search

## Microsoft Corporation

Main site for product information, support, and news....  
[www.microsoft.com/](http://www.microsoft.com/) ---[Business]/[Economy]/[Education]/[Technology]

## Microsoft Download Center

Update for Microsoft Office Outlook 2007 Junk Email Filter (KB2466076) ... Try Microsoft Office 2010 free . Take a free test drive with Windows 7 ...  
[www.microsoft.com/downloads/en/default.aspx](http://www.microsoft.com/downloads/en/default.aspx) ---[Technology]

## Virus, Spyware & Malware Protection | Microsoft Security Essentials

Microsoft Security Essentials provides real-time protection for your home PC that guards against viruses, spyware, and other malicious software. ...  
[www.microsoft.com/security\\_essentials/](http://www.microsoft.com/security_essentials/) ---[Technology]

## Office – Microsoft Office

Try or buy Office 2010 . View product information, get help and training, explore templates, images, and downloads.  
[office.microsoft.com/](http://office.microsoft.com/) ---[Education]/[Technology]

## Microsoft – Wikipedia, the free encyclopedia

Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses, ...  
[en.wikipedia.org/wiki/Microsoft](http://en.wikipedia.org/wiki/Microsoft) ---[Education]/[Technology]

## Education Overview – Resources

Through Partners in Learning, Microsoft is working with education and government ... Gates says Microsoft forced to look for talent in developing countries ...  
<https://65.55.21.250/canada/government/education/default.mspx> ---[Education]/[Society]/[Technology]

## MSDN | Microsoft Development, Subscriptions, Resources, and More

Visit the Microsoft Developer Network to find development resources, training, references, blogs, forums, the MSDN Magazine, MSDN Subscriptions, and more.  
[msdn.microsoft.com/en-us/default](http://msdn.microsoft.com/en-us/default) ---[Education]/[Technology]

## Microsoft Corporation: NASDAQ:MSFT quotes & news – Google Finance

Get detailed financial information on Microsoft Corporation (NASDAQ:MSFT) including real-time stock quotes, historical charts & financial news, ...  
[www.google.com/finance?q=NASDAQ:MSFT](http://www.google.com/finance?q=NASDAQ:MSFT) ---[Business]/[Economy]/[Technology]

## Home : The Official Microsoft Silverlight Site

Microsoft portal site for the Silverlight development community. Download Silverlight, post to the forums, read Silverlight blogs.

Class  
Labels

# Character of our method

Comparing to other methods

- Similar object: layout search results in a more clear and coherent way.
- Different method:
  - Traditional method: clustering
  - **Our method: multi-label classification**
- Advantages:
  - Every class (“cluster”) has a pre-defined name.
  - A text could appear in more than one class.

# Multi-label classification problem

- **Single-label classification:** examples are associated with a single label  $l$  from a set of disjoint labels  $L$ .
- **Multi-label classification:** examples are associated with a set of labels.
- Solving strategies for multi-label classification:
  - **Problem transforming:** transform the multi-label classification problem into one or more single-label classification or regression problems. (adopted in this work)
  - **Algorithm adaptation:** extend specific learning algorithms in order to handle multi-label data directly.

# Transforming method in this work

Ex.	Sports	Religion	Science	Politics
1	X			X
2			X	X
3	X			
4		X	X	



Ex.	Sports	$\neg$ Sports
1	X	
2		X
3	X	
4		X

Ex.	Politics	$\neg$ Politics
1	X	
2	X	
3		X
4		X

Ex.	Religion	$\neg$ Religion
1	X	
2	X	
3		X
4		X

Ex.	Science	$\neg$ Science
1		X
2	X	
3		X
4	X	

The most common problem transformation method (dubbed PT4) learns  $|L|$  binary classifiers , one for each different label.

For the classification of a new instance  $x$ , this method outputs a set of labels that is the union of the labels that are output by the  $|L|$  classifiers.

# Bayesian Multi-label Classification

- Classifier design:  
    Naïve Bayesian (NB) multi-label classification algorithm
- Feature selection:  
    Two-step feature selection strategy

# Naïve Bayesian multi-label classifier (NBML)

- Why we choose Naïve Bayesian classifier?

NB algorithm is efficient. For the real-time system, it has a wide application such as in Spam filtering system.

- The principal of NBML classifier:

For the classification of a new document:

- ü Output a set of labels which is the union of the labels that are output by  $|L|$  single classifiers;
- ü For each single label NB classifier, compute conditional probability of document with relate to each class label;
- ü Design a parameter to decide the label set of a document.

# Two-step feature selection strategy

## – Why we conduct two-step feature selection?

A Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. However, in real classification problems, this assumption is usually hard to be satisfied.

## – Two-step feature selection strategy

- Select discriminative features: DF+CHI-2
- Filter irrelevant and redundant features among classes: FCBF

# Select discriminative features

## (1)

### | DF based method to filter rare feature in each class

In corpus D, each text belongs to a label set  $Y$ . Here,  $Y \subseteq C$ ,  $C = \{c_1, c_2, \dots, c_n\}$  is the class set defined before classification.

Relative text frequency is noted as  $Text\_freq\_relative_{ij}$ .

$$Text\_freq\_relative_{ij} = \frac{Text\_freq_{ij}}{N_i}$$

Here,  $N_i$  is the quantity of texts with related to label  $c_i$  in training set.  $Text\_freq_{ij}$  is the number of texts which include word  $j$  with related to label  $c_i$ .

## Select discriminative features (2)

*Algorithm 1: Filtering rare feature in each class*

For  $c_i \in C$ ,  $C = \{c_1, c_2, \dots, c_i, \dots, c_n\}$ ,

For word  $j \in Term_i$ ,

If ( $Text\_freq\_relative_{ij} < a$ )

{remove word  $j$ ;

Else {word  $j \in Term'_i$ ;

$Term' = \{Term'_1, Term'_2, \dots, Term'_i, \dots, Term'_n\}$

In our work,  $\alpha$  is  
set as 0.02  
according to  
experiments.

$Term_i$  includes all the words extracted in the documents with related to label  $c_i$ ,  $Term'_i$  includes all the words selected in the documents with related to label  $c_i$  and  $Term'$  is the word set in all class labels selected by Algorithm 1

# Select discriminative features

## (3)

- | CHI-2 based method to select discriminative features
  - üCompute CHI-2 value on matrix “feature \* class”
  - üSelect first 5,000 features according to CHI-2 value

# Filter irrelevant and redundant features among classes

- **FCBF algorithm:**  
A fast feature filter method which could identify relevant features as well as redundancy among relevant features.
- [Ref.] Lei Yu, Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

# Classifier evaluation

- **Corpus**
  - Multi-label text classification corpus from NTT communication science research group
- **Metrics**
  - ü Hamming Loss
  - ü One-error
  - ü Coverage
  - ü Average precision

# Classifier evaluation (Hamming Loss)

Text Set	Algorithm			
	NBML	ML-kNN	AdaBoost.MH	Rank-SVM
Arts& Humanities	0.0605	0.0612	<b>0.0585</b>	0.0615
Business& Economy	<b>0.0253</b>	0.0269	0.0279	0.0275
Computers& Internet	<b>0.0373</b>	0.0412	0.0396	0.0392
Education	0.0455	<b>0.0387</b>	0.0423	0.0398
Entertainment	<b>0.0544</b>	0.0604	0.0578	0.0630
Health	0.0433	0.0458	<b>0.0397</b>	0.0423
Recreation& Sports	0.0610	0.0620	<b>0.0584</b>	0.0605
Reference	0.0314	0.0314	<b>0.0293</b>	0.0300
Science	0.0336	<b>0.0325</b>	0.0344	0.0340
Social & Science	0.0241	<b>0.0218</b>	0.0234	0.0242
Society & Culture	0.0553	<b>0.0537</b>	0.0575	0.0555
Average	0.0429	0.0432	<b>0.0426</b>	0.0434

# Classifier evaluation (one-error)

Text Set	Algorithm			
	NBML	ML-kNN	AdaBoost.MH	Rank-SVM
Arts & Humanities	0.5756	0.6330	<b>0.5617</b>	0.6653
Business & Economy	0.1420	<b>0.1213</b>	0.1337	0.1237
Computers & Internet	0.4453	0.4357	0.4613	<b>0.4037</b>
Education	0.5105	0.5207	0.5753	<b>0.4937</b>
Entertainment	0.5209	0.5300	0.4940	<b>0.4933</b>
Health	0.3571	0.4190	0.3470	<b>0.3323</b>
Recreation & Sports	<b>0.5517</b>	0.7057	0.5547	0.5627
Reference	0.4776	0.4730	0.4840	<b>0.4323</b>
Science	0.5731	0.5810	0.6170	<b>0.5523</b>
Social & Science	0.3572	<b>0.3270</b>	0.3600	0.3550
Society & Culture	0.4502	0.4357	0.4845	<b>0.4270</b>
Average	0.4510	0.4711	0.4612	<b>0.4401</b>

# Classifier evaluation (coverage)

Text Set	Algorithm			
	NBML	ML-kNN	AdaBoost.MH	Rank-SVM
Arts & Humanities	5.3471	5.4313	<b>5.1900</b>	9.2723
Business & Economy	2.3210	<b>2.1840</b>	2.4730	3.3637
Computers & Internet	<b>4.4279</b>	4.4117	4.4747	8.7910
Education	3.6840	<b>3.4973</b>	3.9663	8.9560
Entertainment	3.1015	3.1467	<b>3.0877</b>	6.5210
Health	<b>3.0266</b>	3.3043	3.0843	5.5400
Recreation & Sports	4.3838	5.1010	<b>4.3380</b>	5.6680
Reference	<b>3.2022</b>	3.5420	3.2643	6.9683
Science	6.4053	<b>6.0470</b>	6.6027	12.4010
Social & Science	3.8422	<b>3.0340</b>	3.4820	8.2177
Society & Culture	5.8794	5.3653	<b>4.9545</b>	6.8837
Average	4.1474	4.0968	<b>4.0834</b>	7.5075

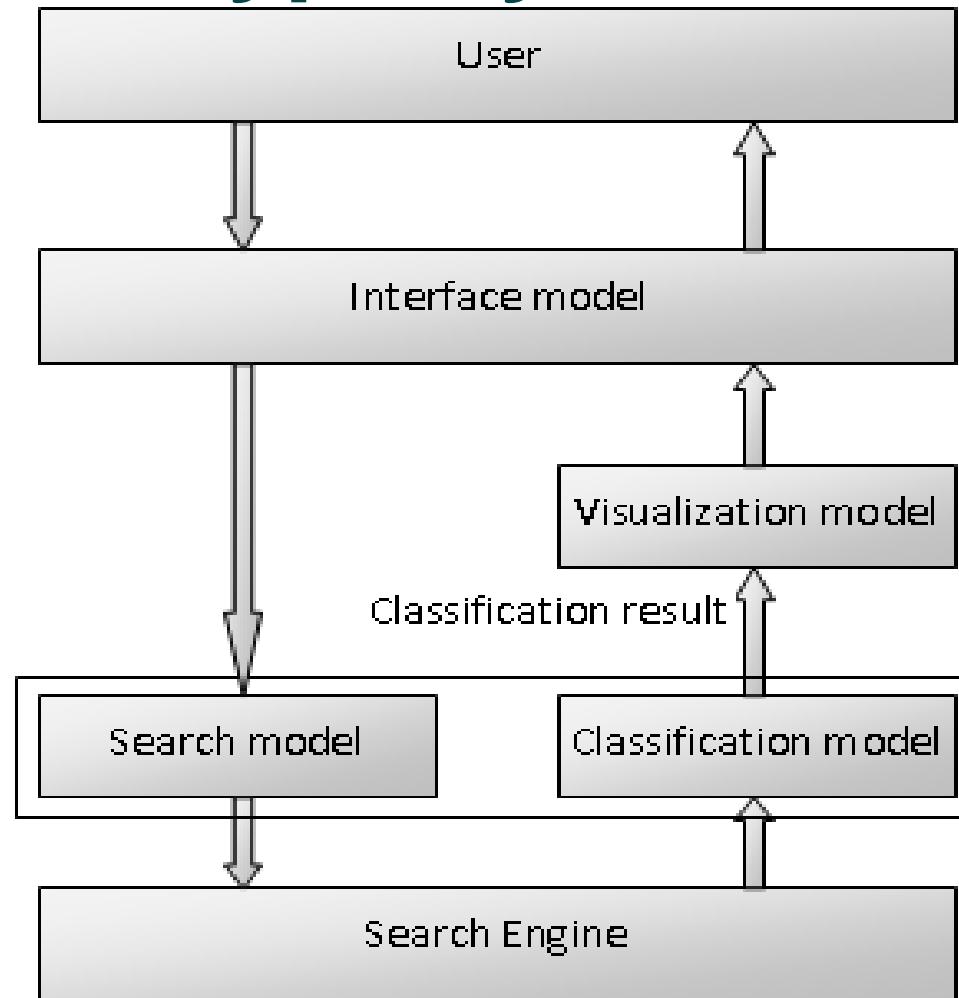
# Classifier evaluation (average precision)

Text Set	Algorithm			
	NBML	ML-kNN	AdaBoost.MH	Rank-SVM
Arts & Humanities	0.5391	0.5097	<b>0.5526</b>	0.4170
Business & Economy	<b>0.8819</b>	0.8798	0.8702	0.8694
Computers & Internet	0.6205	<b>0.6338</b>	0.6235	0.6123
Education	0.5882	<b>0.5993</b>	0.5619	0.5702
Entertainment	<b>0.6503</b>	0.6013	0.6221	0.5637
Health	<b>0.7355</b>	0.6817	0.7257	0.6839
Recreation & Sports	0.5535	0.4552	<b>0.5639</b>	0.5315
Reference	<b>0.6309</b>	0.6194	0.6264	0.6176
Science	0.5234	<b>0.5324</b>	0.4940	0.5007
Social & Science	0.7309	<b>0.7481</b>	0.7217	0.6788
Society & Culture	0.6057	<b>0.6128</b>	0.5881	0.5717
Average	<b>0.6418</b>	0.6249	0.6318	0.6015

# Experiment result

- NBML algorithm has nearly performance comparing to the famous algorithms (i.e. AdaBoost.MH and ML-kNN) in the application of multi-label text classification.
- We don't compare our algorithm with exist NB multi-label algorithm, (i.e. algorithm of McCallum or Minling Zhang) because they have high computational cost. They couldn't satisfy the requirement of real-time text classification system.

# A prototype system: TJ-MLWC



# Testing example of TJ-MLWC

## (1)

Search results of “Microsoft” when selecting “Technology” label

The screenshot shows the TJ-MLWC search interface on the left and the Microsoft search results on the right. On the left, under 'TJ-MLWC' and 'TONGJI UNIVERSITY MULTI-LABEL WEB CLASSIFIER', there is a sidebar with a list of categories: Economy, Society, Education, Technology (which is checked), Politics, Military, Sports, Entertainment, and Business. A red box highlights the 'Technology' checkbox. On the right, the Microsoft search results page has a purple background with a search bar at the top. The results are as follows:

- Microsoft Corporation**  
Main site for product information, support, and news...  
[www.microsoft.com/](http://www.microsoft.com/) --- [Business] / [Economy] / [Education] / [Technology]
- Microsoft Download Center**  
Update for Microsoft Office Outlook 2007 Junk Email Filter (KB2466076) ... Try Microsoft Office 2010 Free - Take a free drive with Windows 7 ...  
[www.microsoft.com/downloads/en/default.aspx](http://www.microsoft.com/downloads/en/default.aspx) --- [Technology]
- Virus, Spyware & Malware Protection | Microsoft Security Essentials**  
Microsoft Security Essentials provides real-time protection for your home PC that guards against viruses, spyware, and other malicious software. ...  
[www.microsoft.com/security\\_essentials/](http://www.microsoft.com/security_essentials/) --- [Technology]
- Office – Microsoft Office**  
Try or buy Office 2010, view product information, get help and training, explore templates, images, and downloads.  
[office.microsoft.com/](http://office.microsoft.com/) --- [Education] / [Technology]
- Microsoft – Wikipedia, the free encyclopedia**  
Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses, ...  
[en.wikipedia.org/wiki/Microsoft](http://en.wikipedia.org/wiki/Microsoft) --- [Education] / [Technology]
- Education Overview – Resources**  
Through Partners in Learning, Microsoft is working with education and government ... Gates says Microsoft forced to look for talent in developing countries ...  
[https://65.55.21.250/canada/government/education/default.mspx](http://65.55.21.250/canada/government/education/default.mspx) --- [Education] / [Society] / [Technology]
- MSDN | Microsoft Development, Subscriptions, Resources, and More**  
Visit the Microsoft Developer Network to find development resources, training, references, blogs, forums, the MSDN Magazine, MSDN Subscriptions, and more.  
[msdn.microsoft.com/en-us/default](http://msdn.microsoft.com/en-us/default) --- [Education] / [Technology]
- Home : The Official Microsoft Silverlight Site**  
Microsoft portal site for the Silverlight development community. Download Silverlight, post to the forums, read Silverlight blogs and learn about ...  
[silverlight.net/](http://silverlight.net/) --- [Education] / [Technology]
- Microsoft Research – Turning Ideas into Reality**  
Computer technology research at Microsoft Corporation

A large red starburst graphic with the word "Technology" is overlaid on the Microsoft search results page.

# Testing example of TJ-MLWC (1)

Search results of “Microsoft” when selecting “Education” label

The screenshot shows the TJ-MLWC interface on the left and a search results page for "Microsoft" on the right. On the left, under the 'TJ-MLWC' logo, there is a sidebar with a list of categories: Economy, Society, Education (which is checked), Technology, Politics, Military, Sports, Entertainment, and Business. The 'Education' category is highlighted with a red box. On the right, the search results for "Microsoft" are displayed. A red starburst points to the word "Education" in the search results. Several links are circled in red, including "Microsoft Corporation", "Microsoft Education: Lesson plans, tutorials & education resources", "Office – Microsoft Office", "Teachers – Microsoft Corporation", "Microsoft – Wikipedia, the free encyclopedia", "Education Overview – Resources", "Microsoft Education Labs", "MSDN | Microsoft Development, Subscriptions, Resources, and More", and "Home : The Official Microsoft Silverlight Site".

■ TJ-MLWC  
TONGJI UNIVERSITY  
MULTI-LABEL WEB CLASSIFIER

Economy  
Society  
**Education**  
Technology  
Politics  
Military  
Sports  
Entertainment  
Business

Microsoft

Microsoft Corporation  
Main site for product information, support, and news....  
[www.microsoft.com/](http://www.microsoft.com/) --- [Business]/[Economy]/[Education]/[Technology]

**Microsoft Education: Lesson plans, tutorials & education resources**  
Microsoft Education offers lesson plans, tutorials and in-depth Microsoft product resources, including Office and Windows, for K-12 and higher education.  
[www.microsoft.com/education/default.mspx](http://www.microsoft.com/education/default.mspx) --- [Education]

**Office – Microsoft Office**  
Try or buy Office 2010, view product information, get help and training, explore templates, images, and downloads.  
[office.microsoft.com/](http://office.microsoft.com/) --- [Education]/[Technology]

**Teachers – Microsoft Corporation**  
Windows Academic Program: Core Windows technologies for higher-education ...  
[www.microsoft.com/education/teachers/default.aspx](http://www.microsoft.com/education/teachers/default.aspx) --- [Education]

**Microsoft – Wikipedia, the free encyclopedia**  
Microsoft Corporation is an American public multinational corporation headquartered in Redmond, Washington, USA that develops, manufactures, licenses, ...  
[en.wikipedia.org/wiki/Microsoft](http://en.wikipedia.org/wiki/Microsoft) --- [Education]/[Technology]

**Education Overview – Resources**  
Through Partners in Learning, Microsoft is working with education and government ... Gates says Microsoft forced to look for talent in developing countries ...  
<https://65.55.21.250/canada/government/education/default.mspx> --- [Education]/[Society]/[Technology]

**Microsoft Education Labs**  
Microsoft Education Labs: Exploring the productivity horizon. Try. Experience. Discuss the future of education.  
[www.educationlabs.com/](http://www.educationlabs.com/) --- [Education]

**MSDN | Microsoft Development, Subscriptions, Resources, and More**  
Visit the Microsoft Developer Network to find development resources, training, references, blogs, forums, the MSDN Magazine, MSDN Subscriptions, and more.  
[msdn.microsoft.com/en-us/default](http://msdn.microsoft.com/en-us/default) --- [Education]/[Technology]

**Home : The Official Microsoft Silverlight Site**  
Microsoft portal site for the Silverlight development community. Download Silverlight, post to the forums, read Silverlight blogs and learn about ...  
[silverlight.net/](http://silverlight.net/) --- [Education]/[Technology]

# Testing example of TJ-MLWC

## (1)

Search results of “Microsoft” when selecting  
“Technology” and “Education” label



# Conclusion from testing examples

- Users could browse search results by selecting one or more interested labels.
- One document could relate to more than one class label.
- Class label could reflect the contents of corresponding documents.

# Conclusion and perspective

## – Conclusion:

- Present a Web text search results visualization method based on multi-label classification.
- Present a two-step feature selection algorithm and a Naïve Bayesian multi-label classification algorithm.
- Design a prototype system TJ-MLWC.

## – Perspective:

- Taxonomy is hard to extend automatically with new contents appearing. Further research focuses on semi-supervised NB multi-label classification algorithm which aims at making use of unlabeled on-line data to improve classifier performance.

Thank you!

&

Question?

A Naïve Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Results  
Duoqian Miao et al. Tongji University