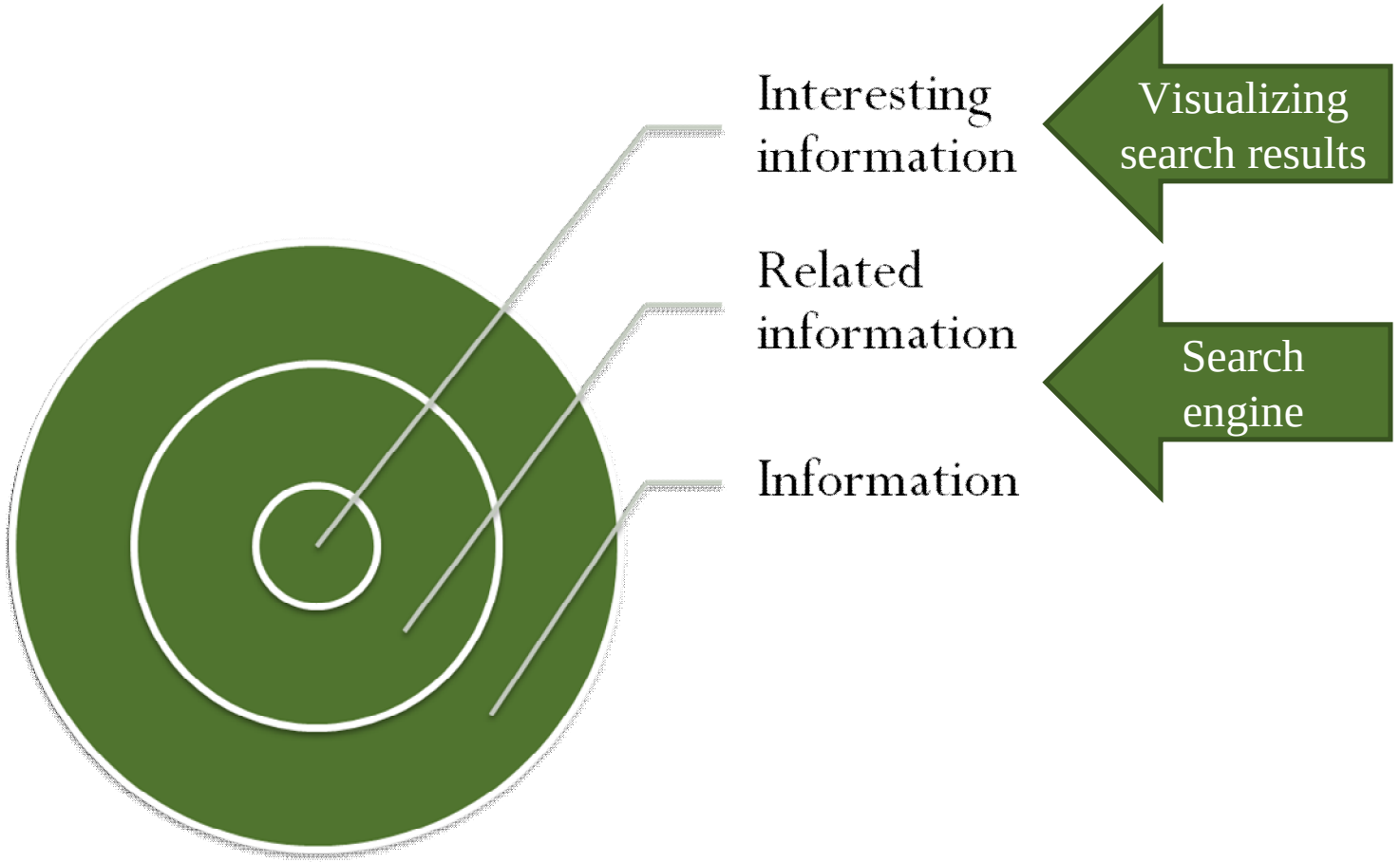
A Naïve Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Results

Duoqian Miao
Tongji University
2011-2-28

Outline

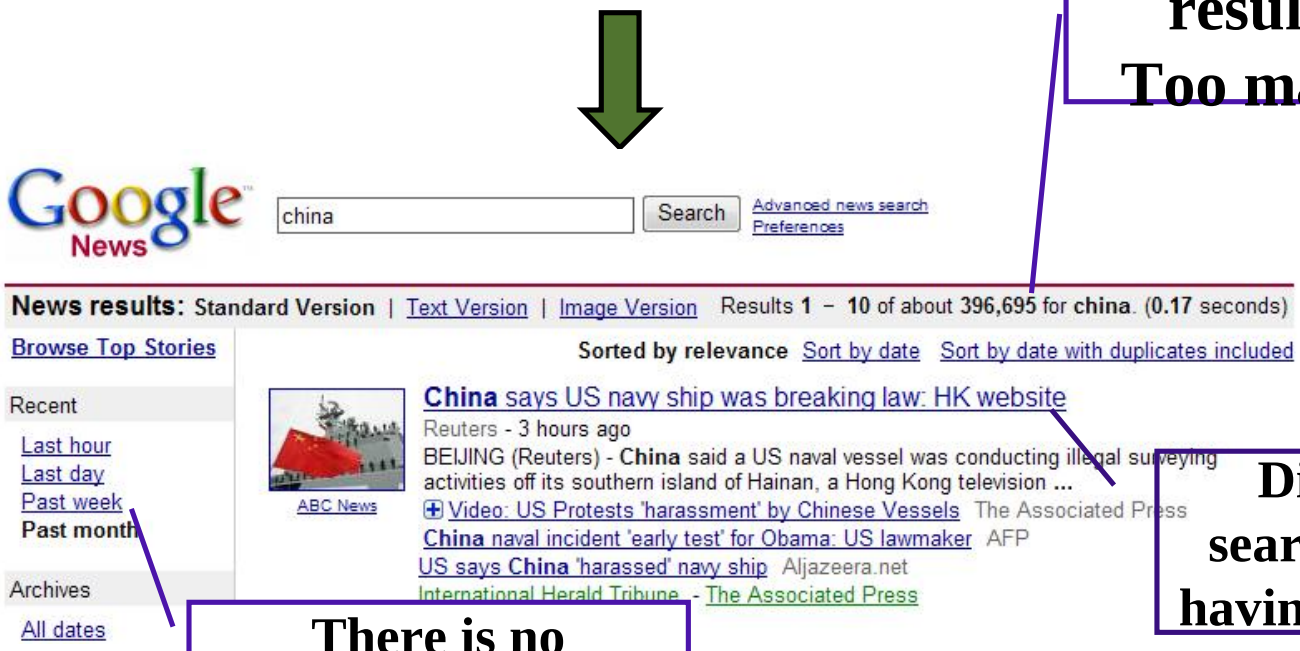
- Why? (Do search results need to be visualized?)
- How? (How could we realize it?)
- Character of our method
- Bayesian multi-label classification algorithm
- A prototype system: TJ-MLWC
- Conclusion

Why?


Interesting information
Related information
Information
Visualizing search results
Search engine

A example: Google News

Search “China” in Google News

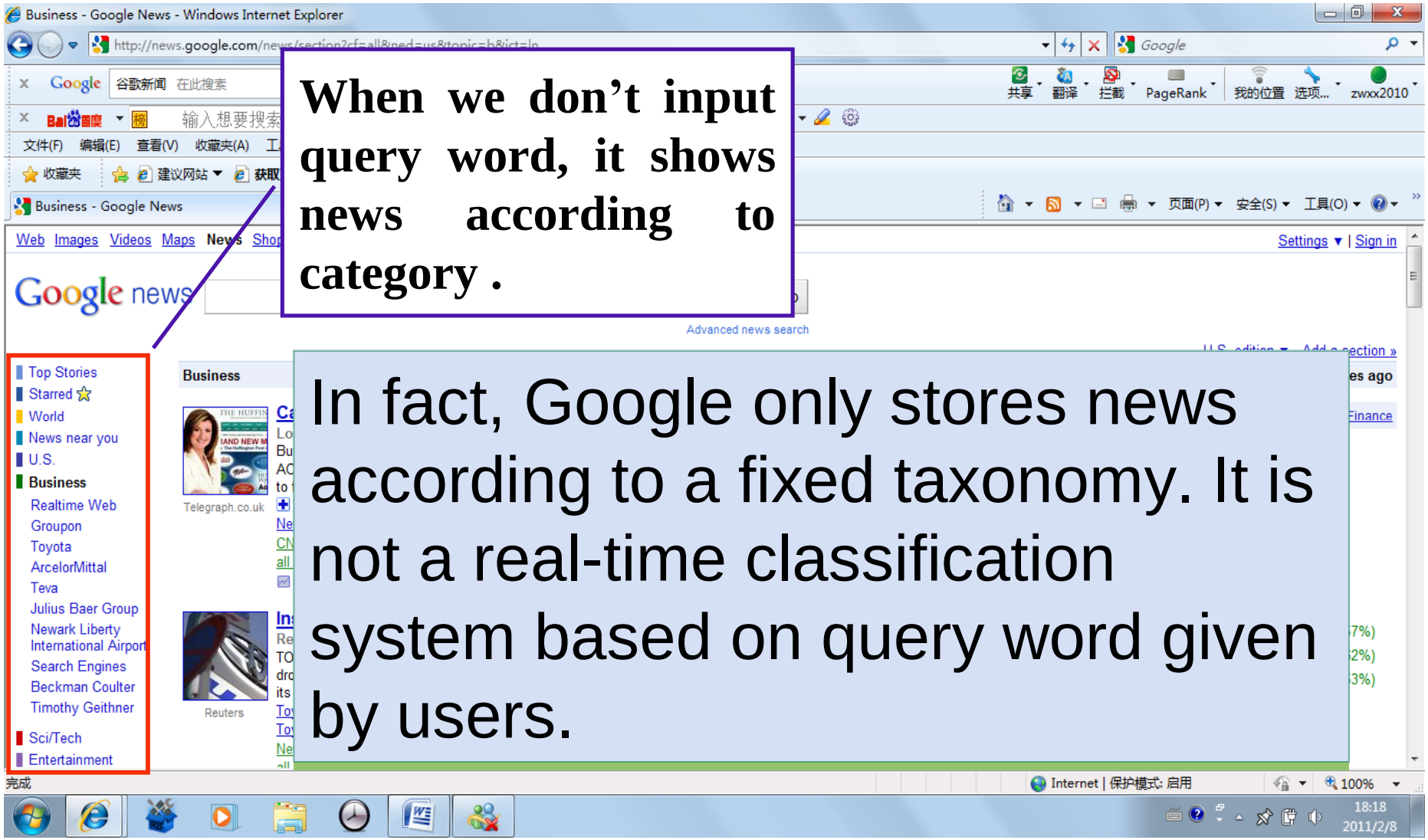


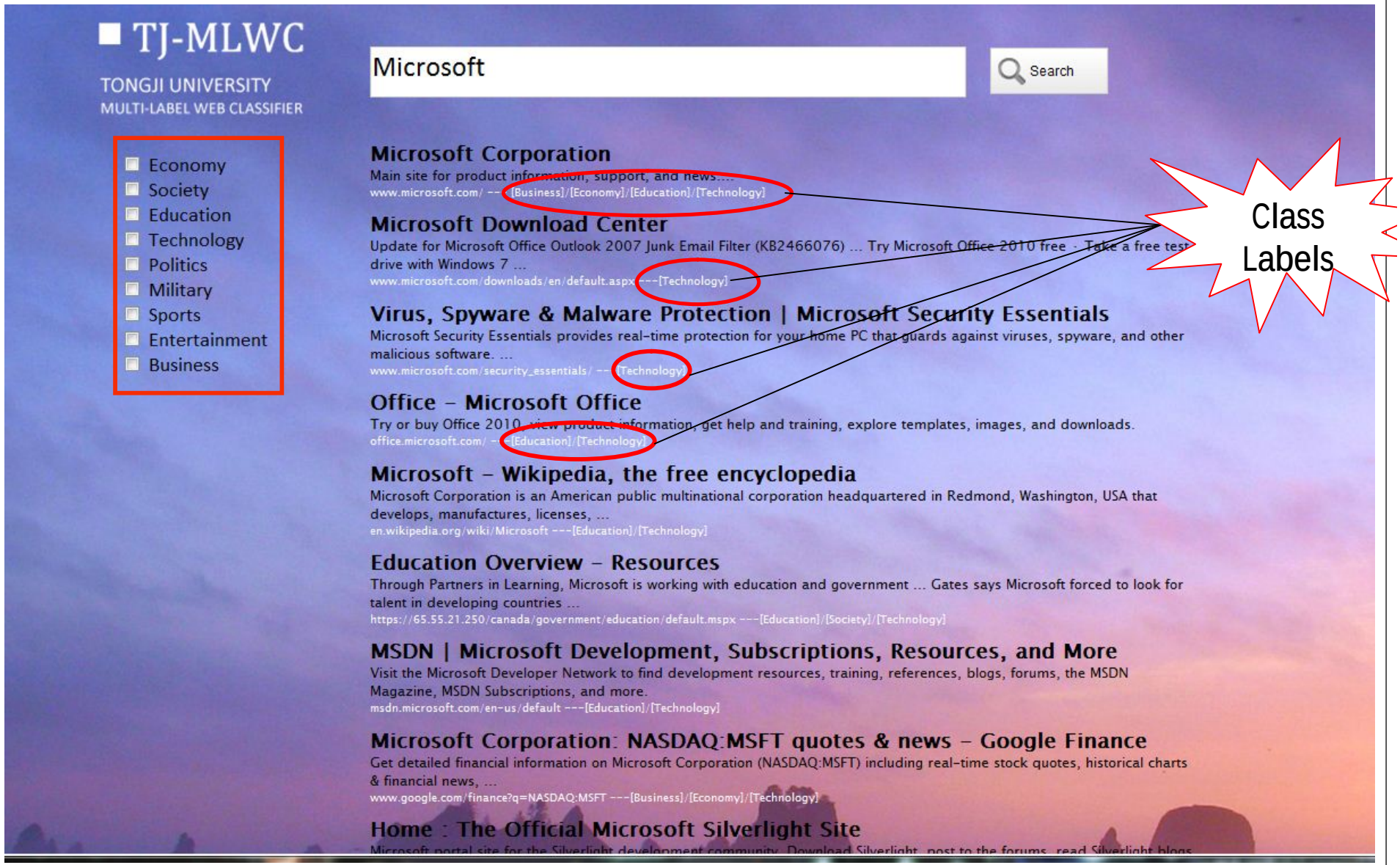
396,695 results! Too many!

Displaying all search results and having no class label.

There is no category with related to query word!

A example: Google News (Cont.)




TJ-MLWC
TONGJI UNIVERSITY
MULTI-LABEL WEB CLASSIFIER
Economy
Society
Education
Technology
Politics
Military
Sports
Entertainment
Business
Microsoft
Search
Microsoft Corporation
Main site for product information, support, and news....
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 ---[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/ ---[Technology]
Office – Microsoft Office
Try or buy Office 2010, view product information, get help and training, explore templates, images, and downloads.
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 ---[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 ---[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 ---[Business]/[Economy]/[Technology]
Home : The Official Microsoft Silverlight Site
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	

PT 4

Ex.	Sports	¬Sports
1	X	
2		X
3	X	
4		X

Ex.	Politics	¬Politics
1	X	
2	X	
3		X
4		X

Ex.	Religion	¬Religion
1	X	
2	X	
3		X
4		X

Ex.	Science	¬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 chooose 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:

  - Output a set of labels which is the union of the labels that are output by |L| single classifiers;
  - For each single label NB classifier, compute conditional probability of document with relate to each class label;
  - 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, $Y \subseteq C$, $C = \{c_1, c_2, ..., c_n\}$ 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 $c_i \in C,\ C = \{c_1, c_2, ...c_i, ..., c_n\}$,

For word $j \in Term_i$,

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

{remove word j;}

Else {word $j \in Term_i^{'}$;}

$Term^{'} = \{Term_1^{'}, Term_2^{'}, ..., Term_i^{'}, ..., Term_n^{'}\}$

In our work, α 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**
  - Compute CHI-2 value on matrix “feature * class”
  - Select 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	**0.0585**	0.0615
Business& Economy	**0.0253**	0.0269	0.0279	0.0275
Computers& Internet	**0.0373**	0.0412	0.0396	0.0392
Education	0.0455	**0.0387**	0.0423	0.0398
Entertainment	**0.0544**	0.0604	0.0578	0.0630
Health	0.0433	0.0458	**0.0397**	0.0423
Recreation& Sports	0.0610	0.0620	**0.0584**	0.0605
Reference	0.0314	0.0314	**0.0293**	0.0300
Science	0.0336	**0.0325**	0.0344	0.0340
Social & Science	0.0241	**0.0218**	0.0234	0.0242
Society & Culture	0.0553	**0.0537**	0.0575	0.0555
Average	0.0429	0.0432	**0.0426**	0.0434

Classifier evaluation (one-error)

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

Classifier evaluation (coverage)

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

Classifier evaluation (average precision)

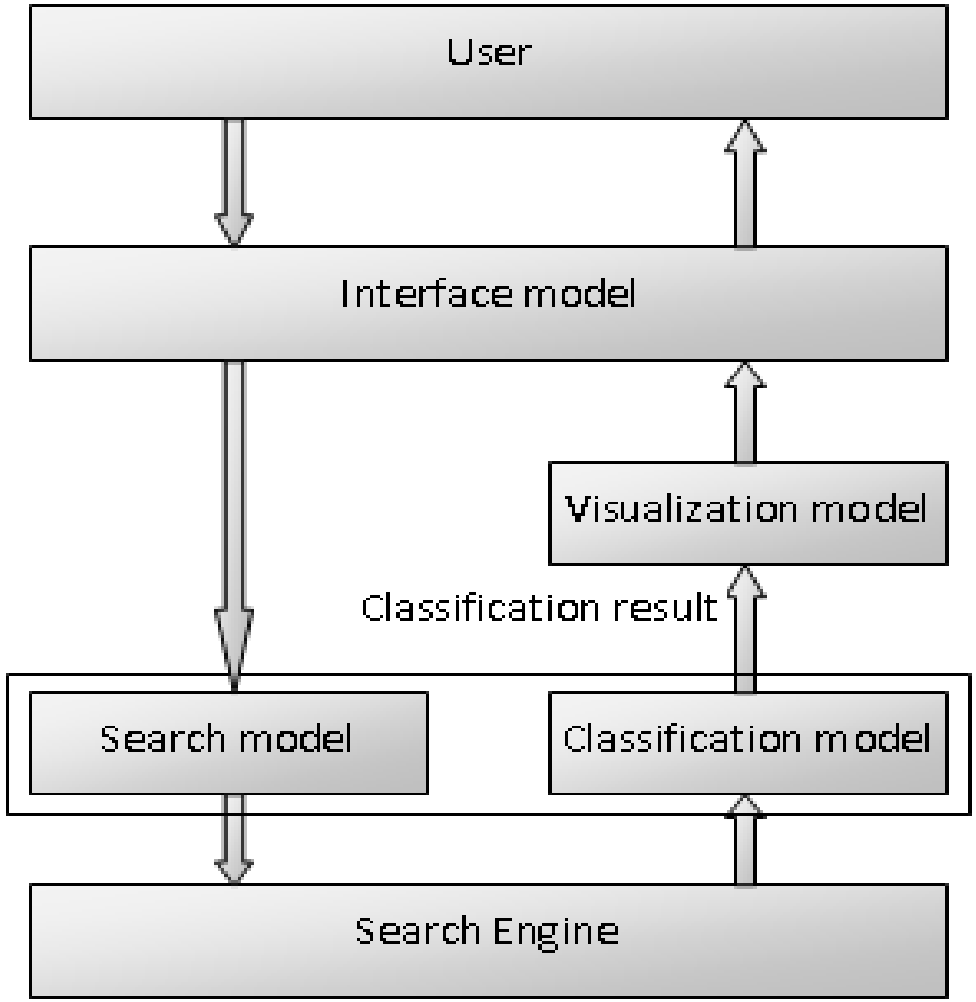
Text Set	Algorithm			
	NBML	ML-kNN	AdaBoost.MH	Rank-SVM
Arts & Humanities	0.5391	0.5097	**0.5526**	0.4170
Business & Economy	**0.8819**	0.8798	0.8702	0.8694
Computers & Internet	0.6205	**0.6338**	0.6235	0.6123
Education	0.5882	**0.5993**	0.5619	0.5702
Entertainment	**0.6503**	0.6013	0.6221	0.5637
Health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

Experiment result

- NBML algorithm has nearly performance comparing to the famous algorithms (i.e. AdaBoost.MH and ML-*k*NN) in the application of multi-label text classification.

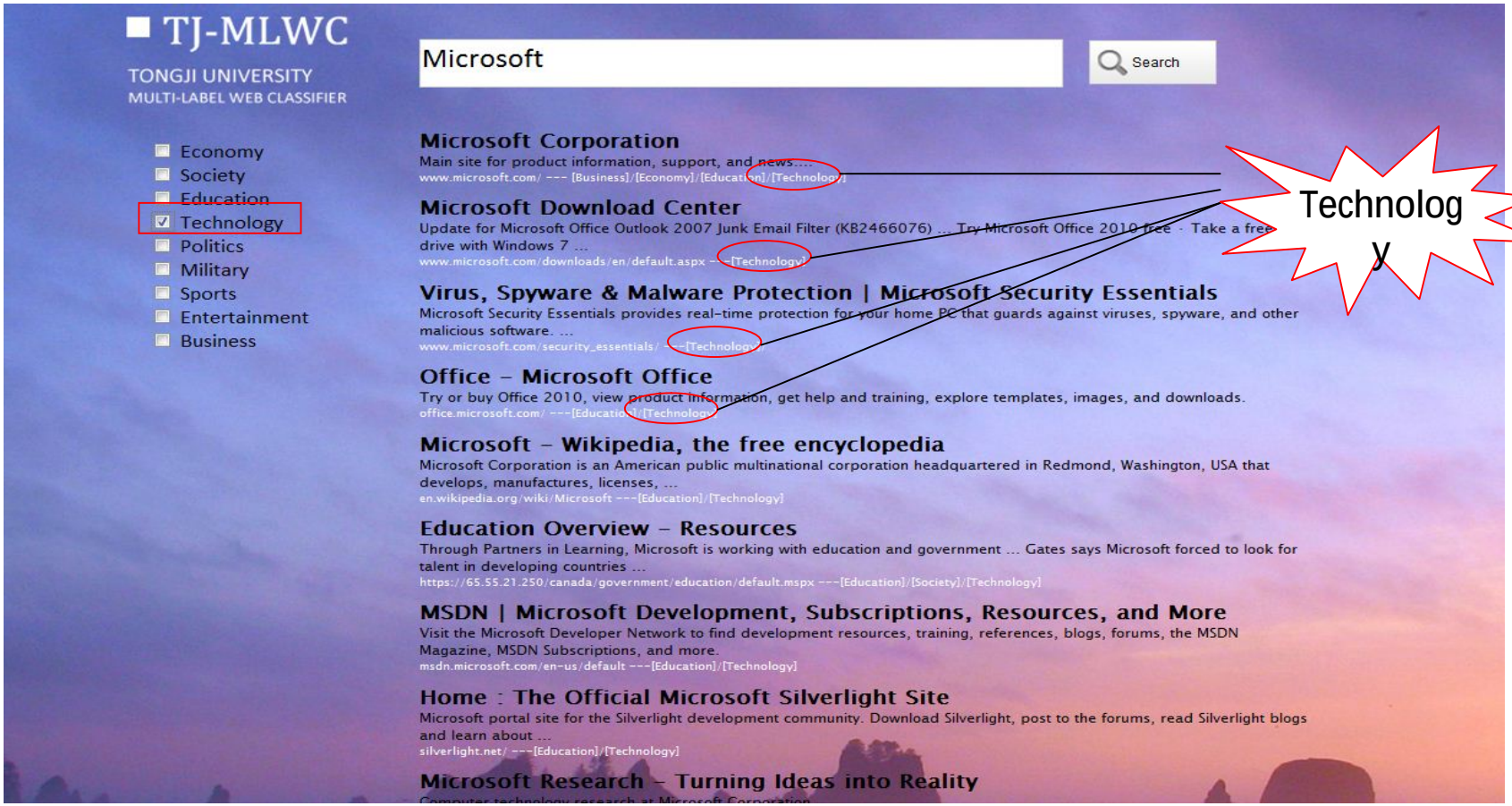
- We don’t compare our algorithm with exist NB multi-label algorithm, (i.e. algorithm of McCallum or Min-ling Zhang) because they have high computational cost. They couldn’t satisfy the requirement of real-time text classification system.

A prototype system: TJ-MLWC



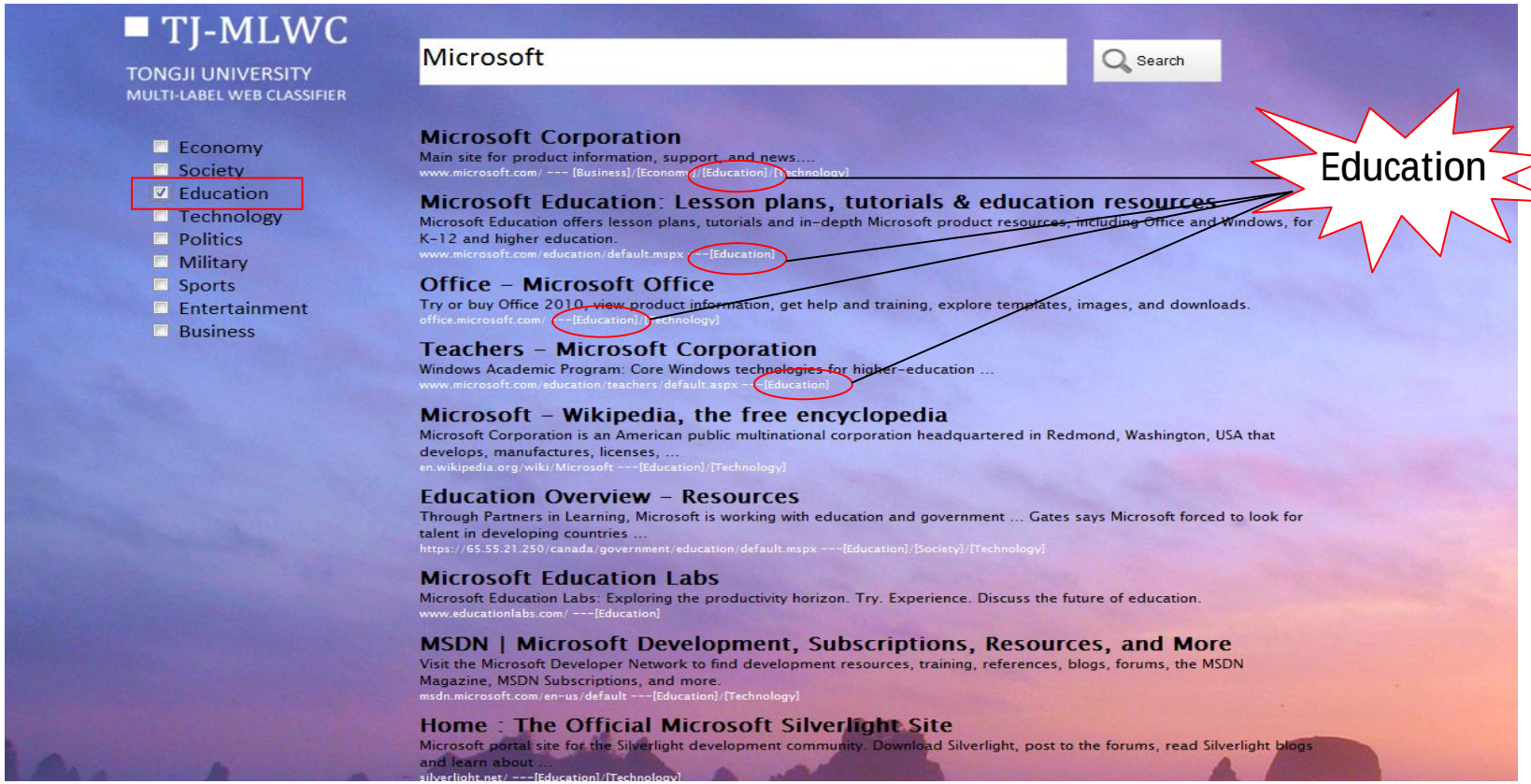
Testing example of TJ-MLWC (1)

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



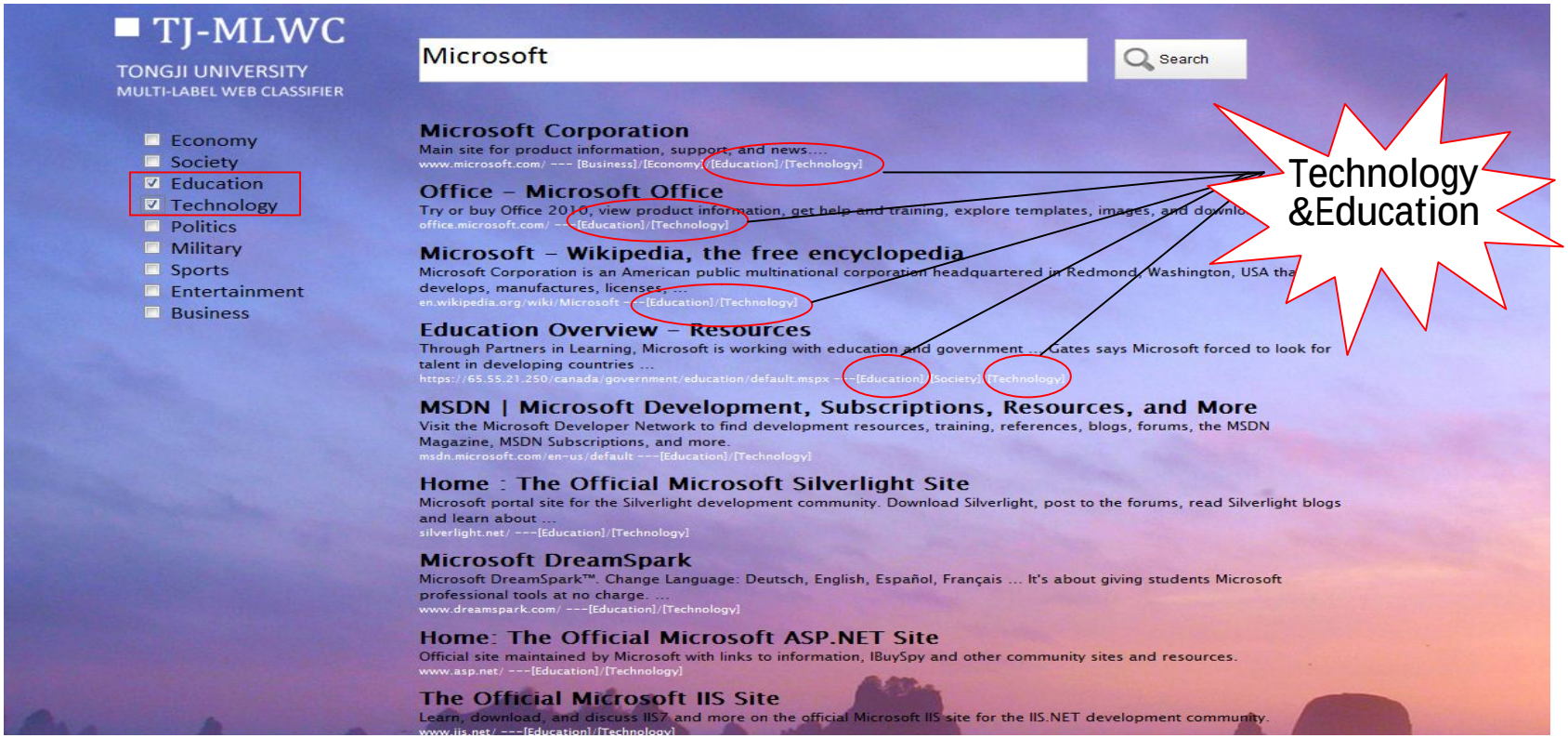
Testing example of TJ-MLWC (1)

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



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