Extracting Knowledge from Community QA (cQA) Sites

Tat-Seng Chua
School of Computing,
National University of Singapore

28 August 2010, COLING Workshop, People’s Web Meets NLP
Information-Rich World

- Web has opened up vast resources of knowledge to us

- **Wikipedia:**
  - Has over 1 million topics, collaboratively authored

- **Yahoo!Answers:**
  - .. and equivalent Q&A sites that provide answers to mostly “how-to” type questions.

- **CNet:**
  - .. and equivalent iCQ sites that provide discussion forums on many topics (in vertical domain)

- **Twitter:**
  - Many live postings on evolving events

- **We can find almost any content on the Web:**
  - From structured knowledge to unstructured information
  - From planned contents to spontaneous deliberations
Two Types of Knowledge Resources

• Community Constructed Semantic Resource (CSR)
  o Contents created collaboratively in structured manner.
  o Has good underlying knowledge structures
  o Example, Wikipedia

• Spontaneous evolving knowledge
  o Spontaneous contributions by users, thru posting questions, expressing opinions and reporting events etc..
  o Represent current topics of interests by users
  o Provide rich sources of timely, mass knowledge
  o Example, cQA sites, forum sites, Twitter etc.
Wikipedia as a Source of Knowledge

• Has been exploited to support various NLP functions
  o From question-answering (various research and POWERSET) to ontology construction, knowledge entailment etc, (this Workshop)

• Limitations of Wikipedia-based QA
  o Effective in supporting a range of factoid and definition–type questions
  o But may be inadequate for how-to, opinion and analysis-type tasks:
    ▪ How-to: “How to transfer pictures from camera to computer?”
    ▪ Opinion-oriented: “Is Kindle a good E-Reader?”
    ▪ Analysis: “Why is iPhone so popular?”
  o Need latest evolving sources of knowledge

• QA is a conscious and goal directed operation
  o But users may not know what they want to ask, they just want to know what are important currently
  o Need to support structured browsing of evolving knowledge
From Unstructured Info to Structured Knowledge

- **AIMS:**
  - To extract structured knowledge from evolving information sources like cQA, forum and live sites
  - To combine with structured knowledge sources like Wikipedia and others to support a wide range of retrieval and analysis functions

- **Focus on 2 knowledge extraction tasks:**
  - From cQA source $\rightarrow$ summary of topic - subtopics
  - From forum source $\rightarrow$ summary of topic and facets with opinion
OUTLINE:

• Finding similar questions in cQA
• Organizing and navigating cQA archives
• Extracting knowledge from forum postings
Answering “How-To” Questions

- Very difficulty – both problems in understanding the question and generating the answers
- Community QA systems are good resource to offer this service
  - Yahoo!Answer has tremendous QA resources in archive: 60M QA pairs in YA (the latest statistic, Dec 2008)
- Simplify problem as a similar question search problem

However, it is still a difficult problem:
1. Long verbose natural language query
2. Noisy and Ungrammatical language
3. Multiple forms of questions
4. Uneven quality of answer

- Multiple forms of questions:
  o How can I lose weight in a few month?
  o What should I do to lose pound in a short period?
Answering “How-To” Questions -2

- For the 1st problem, an example of a long cQA question with multiple sub-questions is:
  (do you have health insurance?)\(^1\) (is it provided by your employer?)\(^2\) (do you supply your own?)\(^3\) (if so how much do you have to pay)\(^4\) and (is it worth it for services you get?)\(^5\)?

- Retrieval results without segmentation:
  - … not effective

<table>
<thead>
<tr>
<th>Rank</th>
<th>Matched Questions</th>
</tr>
</thead>
</table>
| 1    | qid=20071203201922AAamL5K  
How do you pay your health insurance deductibles? |
| 2    | qid=20080528102757AA5qfMK  
How can I get help paying for health insurance? |
| 3    | qid=20080103120157AAjhfwr  
Health Insurance Through Employers? |
| 4    | qid=20080208204910AAC1wWb  
health insurance during a switch of employment? |
| 5    | qid=20080425051938AA8be20  
How much would I pay for a tonsillectomy if I did not have health insurance? |
| 6    | qid=20080416173725AAPDCPI  
how to get health insurance? |
| 7    | qid=20080520055704AAJXnft  
How to find best health insurance provider? |
Answering “How-To” Questions -3

- Question segmentation
  - Allows for better separation & matches for each sub-questions & answers

- Our approach to tackle problem 1 (SIGIR’10, COLING’10):
  - Build a question detector to distinguish between question (Q) & context (C)
  - Measure closeness between Q and C sentences
  - Build Q-C relationship to form a graph
  - Use the graph to propagate the closeness scores
  - Group topically related Q and C sentences

- Example
  - Query: “what are the requirements of being a dentist?”
  - A good matching sub-question (after question segmentation) is:

    i heard somewhere that in order to become a dentist, you need certain hours of volunteering or shadowing.
    is that true? if it is, how many hours? i have only a few hours of 
    and can you write down other requirements that one would need i know there are a lot of things but if you can write down as 
    i can, that’d be a lot of help. thanks

  Lead to 10% Improvement!
Answering “How-To” Questions -4

- To tackle problems 2 & 3 (SIGIR’09):
  - Explore Syntactic Tree Matching
  - Perform tree decomposition
  - Match tree fragments at production rule level, more robust against grammatical errors

- Parsing Tree for the sentence: “Good workout plan for losing a little bit of weight?”

- Parsing Tree for another variant of sentence: “Good workout plan to losing a little bit of weight?”

- Multiple forms of questions:
  - How can I lose weight in a few month?
  - What should I do to lose pound in a short period?

- Structures of sentence at production rule level remain the same!!
Answering “How-To” Questions

- Perform question and answer segmentation
- Semantic-smoothed Tree Matching
  - Perform semantic similarity matching between leaf words
- Perform Answer Matching
  - Questions with similar answers tend to be semantically similar too
- Overall Retrievval Model

Diagram:

- Query
- Top k
- Syntactic Tree Matching
- Result 2
- Final Result
- Lucene
- Q Index
- Result 1
- A Index
- Result
- Segmented Questions
- Segmented Answers
- Yahoo! Answer
**Result Analysis - Accuracy**

**Query question**: what is the reason that my gums bleeding when I brush my teeth?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Matched Questions</th>
<th>Results returned by BoW method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>qid=20080330042256AAYGcpJ bleeding gums whenever brush teeth?</td>
<td>qid=20080113135256AAAX72kk My gums bleed when i brush my teeth, why?</td>
</tr>
<tr>
<td>2</td>
<td>qid=20080102233315AArC6GR What are the reasons my gums could be bleeding?</td>
<td>qid=20080221211719AAAPMD0 why do my gums bleed when i brush my teeth??</td>
</tr>
<tr>
<td>3</td>
<td>qid=20080505223640AAPMuYc what is the reason for bleeding in gums?</td>
<td>qid=20080211211719AAAPMD0 why are my gums bleeding when I brush my teeth?</td>
</tr>
<tr>
<td>4</td>
<td>qid=20080221211719AAAPMD0 why do my gums bleed when i brush my teeth??</td>
<td>qid=20080211211719AAAPMD0 why are my gums bleeding when I brush my teeth?</td>
</tr>
<tr>
<td>5</td>
<td>qid=2008050503109AAKs4Ba why are my gums bleeding when I brush my teeth?</td>
<td>qid=20080211211719AAAPMD0 why are my gums bleeding when I brush my teeth?</td>
</tr>
<tr>
<td>6</td>
<td>qid=20080113135256AAAX72kk My gums bleed when i brush my teeth, why?</td>
<td>qid=20080505223640AAPMuYc what is the reason for bleeding in gums?</td>
</tr>
<tr>
<td>7</td>
<td>qid=20080518173957AAZIYkN whenever i brush my teeth, my gums start to bleed, is that a problem, and how do i stop it?</td>
<td>qid=20080326043133AAAX74HO Do your gums bleed when you wash your teeth?</td>
</tr>
<tr>
<td>8</td>
<td>qid=20080107034333AAC7hFl I was brushing my teeth and my gums started bleeding...</td>
<td>qid=20080330042256AAYGcpJ bleeding gums whenever brush teeth?</td>
</tr>
<tr>
<td>9</td>
<td>qid=20080317152130AAQLPw4 My gums bleed when i brush my teeth.. not all the time but quite often..?</td>
<td>qid=20080113135256AAAX72kk My gums bleed when i brush my teeth, why?</td>
</tr>
<tr>
<td>10</td>
<td>qid=20080506115638AAHgc62 When I brush my teeth, my gums bleed... at first it was a little but now its more...?</td>
<td>qid=20080105084121AAS7kYG my teeth bleed when i brush them?</td>
</tr>
</tbody>
</table>

**Results returned by STM method**
Result Analysis – robustness

Original Query: what is the reason that my gums bleeding when I brush my teeth?
Query with grammatical errors: wat the reazon my gum bleding when I brush my teeth?

Original result list

<table>
<thead>
<tr>
<th>Rank</th>
<th>Matched Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>qid=20080113135256AAAX72kx My gums bleed when i brush my teeth, why?</td>
</tr>
<tr>
<td>2</td>
<td>qid=20080221211719AA5PMD0 why do my gums bleed when i brush my teeth??</td>
</tr>
<tr>
<td>3</td>
<td>qid=2008050503109AAKs4Ba why are my gums bleeding when I brush my teeth?</td>
</tr>
<tr>
<td>4</td>
<td>qid=20080102233315AARc6GR What are the reasons my gums could be bleeding?</td>
</tr>
<tr>
<td>5</td>
<td>qid=2008050523640AAPMuYc what is the reason for bleeding in gums?</td>
</tr>
<tr>
<td>6</td>
<td>qid=20080326043133AAx74HO Do your gums bleed when you wash your teeth?</td>
</tr>
<tr>
<td>7</td>
<td>qid=2008030042256AAAYGcpJ bleeding gums whenever brush teeth?</td>
</tr>
<tr>
<td>8</td>
<td>qid=20080518173957AAZiYkn whenever i brush my teeth, my gums a problem, and how do</td>
</tr>
<tr>
<td>9</td>
<td>qid=2008031752130AAQLpW4 My gums bleed when i brush but quite often..?</td>
</tr>
<tr>
<td>10</td>
<td>qid=20080105084121AAS7kYG my teeth bleed when i brush them?</td>
</tr>
</tbody>
</table>

New result list (with errors)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Matched Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>qid=20080113135256AAAX72kx My gums bleed when i brush my teeth, why?</td>
</tr>
<tr>
<td>2</td>
<td>qid=20080221211719AA5PMD0 why do my gums bleed when i brush my teeth??</td>
</tr>
<tr>
<td>3</td>
<td>qid=2008050503109AAKs4Ba why are my gums bleeding when I brush my teeth?</td>
</tr>
<tr>
<td>4</td>
<td>qid=20080102233315AARc6GR What are the reasons my gums could be bleeding?</td>
</tr>
<tr>
<td>5</td>
<td>qid=2008050523640AAPMuYc what is the reason for bleeding in gums?</td>
</tr>
<tr>
<td>6</td>
<td>qid=2008050821356AA6ZjWJ Is Chewing Gum The Same As Brushing Your Teeth?</td>
</tr>
<tr>
<td>7</td>
<td>qid=20080326043133AAx74HO Do your gums bleed when you wash your teeth?</td>
</tr>
<tr>
<td>8</td>
<td>qid=2008030042256AAAYGcpJ bleeding gums whenever brush teeth?</td>
</tr>
<tr>
<td>9</td>
<td>qid=2008031752130AAQLpW4 My gums bleed when i brush but quite often..?</td>
</tr>
<tr>
<td>10</td>
<td>qid=20080319073158AAlZWpL why are the top of my gums red and bleed when i brush my teeth?</td>
</tr>
</tbody>
</table>

Less affected! Overall improvement in results is 12%
OUTLINE:

• Finding similar questions in cQA
• Organizing and navigating cQA archives
• Extracting knowledge from forum postings
Extracting Structured Knowledge from cQA

• Ability to find previously asked questions with readily answers is useful
  o But it is insufficient, as users may not know what to ask
  o They may want to know what important questions have been asked, from which to learn about the key characteristics and issues of the topic/product

• Hence need to organize all related cQA questions around a topic
  o Classic problem of extracting knowledge from informal info source
  o cQA questions represent the current topics of interests by users
  o Combining with established knowledge structures from product sites or Wikipedia, allows knowledge to be evolving and self-updating
Prototype Hierarchy-based Clustering (PHC) approach to organizing cQA archives (SIGIR’10)
Yahoo! Answers: Query topic “Dental”

- Found abundant set of questions, with best answer for each question
- Disorganized, with no topic-wise structure to help in topic navigation
Where to obtain the initial prototype Hierarchy? From Wikipedia Categories or Product Web Sites..
Wikipedia Pages

• The content of a Wikipedia page as the topic description
The phenomenon of gemination arises when two teeth develop from one tooth bud and, as a result, the patient has a larger tooth but a normal number of teeth overall...
Another example of prototype hierarchy formed from Wikipedia article, product pages, and product user manual.
PHC approach to organizing cQA archives
PHC: A multi-criterion optimization model

1) Similarity between objects are measured by appropriate distance metrics, so as to partition objects into homogeneous clusters that are similar to each other.

2) The data hierarchy and prototype hierarchy are adequately matched, both the nodes and relations, with mismatching handled.

3) The data hierarchy, like a taxonomy or an ontology, is incrementally evolving into a compact structure encoding underlying topics of the collection.
PHC: Four Objective Functions

- **Obj1: Minimum Evolution**
  The optimal hierarchy organizes the whole collection with least information changes to the initial data hierarchy

- **Obj2: Prototype Centrality**
  It assumes that a prototype is located at the center of an object cluster in the object space

- **Obj3: Prototype-Data Hierarchy Resemblance**
  It considers the resemblance of the common hierarchies of the data hierarchy and the prototype hierarchy

- **Obj4: Category Cohesiveness**
  The collection is categorized such that objects in same category are similar to each other and those in different categories are dissimilar to each other
Similarity between Questions

- To accurately measure the similarity score between questions for the objective functions, we adopt two state-of-the-art question retrieval models and their combinations:
  - Translation-based Language Model (TBLM) [Xue et al, SIGIR08]
  - Our Syntactic Tree Kernel Matching Model (STKM) [Wang et al. SIGIR09]
  - Linearly combining them an enhanced measure (TBLM-STKM)
PHC:
Assigning QA pairs to leaf prototype nodes

Yahoo!Answer question on “Dental”
1. What are the symptoms of a reininfected root canal?
2. can you still have a wisdom tooth eruption when your 25 years old or older?
3. How do I strengthen my weak dental Enamel, other than the regular brushing and flossing?
4. What is the most effective mouthwash if you are very sensitive to burning?
5. What are some good toothpaste to ease the pain and whiten my teeth from braces?
6. Why does my jaw click when I open and close my mouth, is that normal?
7. What kind of mouthwash should I use to get rid of gum disease?
8. How bad is lemon juice for your tooth enamel?
9. What kind of mouthwash should I use to get rid of gum disease?
10. How many times should you floss each day without damaging your gums?
...
Dental Floss

- How often should you floss if you have braces?
- How many times should you floss each day without damaging your gums?
- What part of your dental routine do you floss? Before brushing but after listerine?
- what is the best way to floss a dental bridge?
- ...

Tooth Enamel

- What are the cosmetic options available for weak dental enamel?
- Does dental enamel layer re-grow after the layer gets reduced by rough brushing?
- Can dental work wear enamel off your teeth?
- Is dental scaling bad on teeth enamel?
- How do I strengthen my weak dental enamel, other than the regular brushing and flossing?
- ...

PHC:
Some Examples of sub-Topics and FAQs
Similarly for **iPhone**:
Assigning QA pairs to leaf prototype nodes

1. Will the iphone (using MobileMe) enable all my emails (6 business accounts) to be pushed instantly to the phone?
2. How do you save a photo from iPhone safari browser?
3. Can we download pictures using safari in the iphone?
4. How can view my safari history on my iPhone?
5. Does anyone know a mixtape website that works on the iPhone safari browser?
6. How do you transfer ringtones and music from itune to your iphone?
7. Can I still use Iphone apps store and Itune after jailbreaking iphone with blacksn0w?
8. Does iTune Movies Have Subtitles?
9. Does the Compass map thing work with iphone 3gs in sweden?
10. Rented itune movies: how long do they last on your computer or iphone?
## Dataset for Preliminary Testing

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Dental</th>
<th>Ipod</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prototype Hierarchy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max. Depth</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>#. Concepts</td>
<td>104</td>
<td>87</td>
</tr>
<tr>
<td>#. Leaf concepts</td>
<td>62</td>
<td>51</td>
</tr>
<tr>
<td><strong>Collection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#. cQA pairs</td>
<td>6735</td>
<td>4381</td>
</tr>
</tbody>
</table>
Comparison Systems

**Baselines (Unsupervised):**

1) **proKmeans:**
   - prototype hierarchy enhanced K-means divisive hierarchical clustering.

2) **LiveClassifier [Huang et al. WWW04]:**
   - a state-of-the-art hierarchical classifier.

**Supervised:**

1) **CFC Classifier [Guan et al. WWW09]:**
   - a state-of-the-art supervised text categorization technique
# Preliminary Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Dental</th>
<th>Ipod</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Micro-F1</td>
<td>Macro-F1</td>
</tr>
<tr>
<td>proKmeans (baseline1)</td>
<td>0.601</td>
<td>0.592</td>
</tr>
<tr>
<td>LiveClassifier (baseline2)</td>
<td>0.683</td>
<td>0.663</td>
</tr>
<tr>
<td>PHC-BOW</td>
<td>0.755</td>
<td>0.703</td>
</tr>
<tr>
<td>PHC-TBLM</td>
<td>0.803</td>
<td>0.783</td>
</tr>
<tr>
<td>PHC-STKM</td>
<td>0.764</td>
<td>0.721</td>
</tr>
<tr>
<td>PHC-TBLM-STKM</td>
<td><strong>0.885</strong></td>
<td><strong>0.879</strong></td>
</tr>
<tr>
<td>Supervised CFC</td>
<td>0.879</td>
<td>0.904</td>
</tr>
</tbody>
</table>
Summary

• We propose a prototype-hierarchy based clustering (PHC) algorithm, that utilizes both the initial prototype hierarchy and unstructured contents, to structure information
  o Perform clustering based on a multi-criterion optimization function
  o Construct a hierarchy of data clusters from the view of taxonomy structure evolution
  o Achieve better performance than a strong baseline and a state-of-the-art system
• Resulting structures provide evolving knowledge for users
• We plan to improve the efficiency of PHC algorithm and explore its applicability to wider domains
  o Example, to select or compose the best answer for each question cluster
OUTLINE:

• Finding similar questions in cQA
• Organizing and navigating cQA archives
• Extracting knowledge from forum postings
Forum and Sentiment Analysis

• Rich opinion contents
  o News posting, forum, Blog, QQ posting
    ▪ >1.3 million BBS (Up to 2008 in China)
    ▪ >3 billion blogs of 0.181 billion user, >1.131 billion updating/per month (Up to 2009 in China)
    ▪ >2 million postings/per day in Baidu bar
    ▪ >50 million QQ groups with million opinion postings /per minute

• Goal of sentiment analysis
  o To explore the hot topics that people are talking about and their sentiments
  o To provide overview to a range of users – from Government, Business to End-Users
Extracting Structured Knowledge from Forum Postings

• Needs structured overview of aspects of any topic that people are concerned with
Structured Knowledge from Forum Postings
Short Text Sentiment Classification

- Sentiment of a posting is not only determined by the presence of terms, but also by their ordering

<table>
<thead>
<tr>
<th>Text</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>drinks were <strong>expensive</strong> but very <strong>good</strong></td>
<td>Sentiment It</td>
</tr>
<tr>
<td>drinks were <strong>good</strong> but very <strong>expensive</strong></td>
<td>Sentiment It</td>
</tr>
<tr>
<td>the concierge was rather <strong>imperious</strong> instead of being <strong>friendly</strong></td>
<td>Sentiment It</td>
</tr>
<tr>
<td>the concierge was rather <strong>friendly</strong> instead of being <strong>imperious</strong></td>
<td>Sentiment It</td>
</tr>
<tr>
<td>some movies are like a <strong>tasty</strong>; this one is a <strong>feast</strong></td>
<td>Sentiment It</td>
</tr>
<tr>
<td>some movies are like a <strong>feast</strong>; this one is a <strong>tasty</strong></td>
<td>Sentiment It</td>
</tr>
</tbody>
</table>

- Key technologies: Term order sensitive model, semantic smoothing
Structured Knowledge from Forum Postings
Mining the Structure of Topic “Hotel”

- Mine important aspects of a topic with related sentiment in the reviews and organize them using a hierarchy.
- Start with initial topic hierarchy extracted from the Web:
  - Perform aspect identification
    - For example, identify the aspect “price” in review “it’s so expensive, how?”
    - Extraction vs. Clustering
  - Sentiment classification
  - Hierarchy generation
- From the hierarchy tree, we can answer questions like:
  - Which aspects in a specific hotel that people are concerned with, and what are their feelings towards these aspects?
Structured Knowledge from Forum Postings
Mining the Structure of Topic “iPhone”

- Initial topic hierarchy: iPhone 3GS hierarchy
Structured Knowledge from Forum Postings
Ranking of Aspects in Reviews

- People comment on various aspects in the reviews
  - But different aspect has different impact on the overall sentiment
  - We can rank the aspect according to the degree of the impact
  - For example, the important aspects which influence the people’s feeling of a hotel are “location” and “price”

- Hence we can input the reviews of a specific hotel, and obtain the important aspects that people are most concerned with from the reviews
Conclusion

• This talk focuses on:
  o Extracting structured knowledge from evolving information sources like cQA, forum and live sites
  o Combining with structured knowledge sources like Wikipedia and others to support a wide range of retrieval and analysis functions

• Demonstrated on 2 knowledge extraction tasks:
  o From cQA source such as the Yahoo!Answers
    ➔ Structured summary of topic
  o From forum source
    ➔ Structured summary of topic and facets with opinion

• Summary provides structured knowledge to facilitate browsing and learning by users

• Work is only beginning and evolving
Analysis on multimedia resources

- Will extend techniques to vertical domain such as those from Cnet
- ... and extend to multilingual and multimedia resources
  - Find Chinese answers from say Baidu-I-Know
  - Find media answers from Flickr, Youtube

For example:
  - How to transfer photos from camera to computer?
  - How to check engine oil?

- Beyond text – Sentiment media search
  - What about the people opinion expressed in and towards these video/image/audio?
Question-Answer?

Contact: Chua Tat-Seng (chuats@comp.nus.edu.sg)
Towards New Search Paradigm on Web Scale Info Collections -2

- Overall Goal, \( \text{WebQA} = w\text{QA} + c\text{QA} + v\text{QA} + i\text{QA} \)
  - Wikipedia (w), Yahoo!Answers (c), YouTube (v), Flickr (image); and equivalent in Chinese

- Target: Multilingual & Multimedia web-QA