Social Network Analysis

bouguessa.mohamed@uqam.ca
Web today
Web today – Diverse applications

Diverse on-line computing applications
Web today – Millions of users
Web today – Rich content
Web today – Highly dynamic

Content is constantly being updated and changed
Web today – Traces of activity

Massive traces of human social activity are collected
Web today – Rich interactions

Rich interactions between users and content
Web today – social networks
Six degrees of separation

We can all be connected through a series of six contacts appeals to me. It makes the world seem less brutal, and more warm and more friendly.
Why study networks?

• **Build understanding and theory:**
  – How users create content and interact with it and among themselves?

• **Build better on-line applications:**
  – How to design better services and algorithms?
Social Networks Analysis

• A social network is a social structure of people, related (directly or indirectly) to each other through a common relation or interest.

• Social network analysis (SNA) is the study of social networks to understand their structure and behavior.
Social Networks

- Social network: relationship among interacting units.
Social Networks

Interacting unites: Actors / nodes

discrete individual, corporate, or collective social units
Relational ties between actors are channels to transfer, exchange or flow of resources.

Social Networks

Relations, linkages or ties
**Social Networks**

- Social network representation
  - Adjacency matrix (socio-matrix)
  - Graph (Socio-graph)
Key Drivers for CS Research in SNA

• Computer Science has created the cyber infrastructure for
  – Social Interaction
  – Knowledge Exchange
  – Knowledge Discovery

• Ability to capture
  – different about various types of social interactions
  – at a very fine granularity
  – with practically no reporting bias

Data mining techniques can be used for building
descriptive and predictive models of social interactions
SNA Techniques

Prominent problems

• Social network extraction/construction
• Identifying prominent/trusted/expert actors
• Identifying Spammers
• Discovering communities in social networks
• Evolution of social networks
• Link prediction
• Approximating large social networks
Social Network Extraction

• Mining a social network from data sources
• Recent research suggest that there are three sources of social network data on the web
  • Content available on web pages (e.g. user homepages, message threads etc.)
  • User interaction logs (e.g. email and messenger chat logs)
  • Social interaction information provided by users (e.g. social network service websites such as Orkut, Friendster and MySpace)
SNA Techniques

Prominent problems

• Social network extraction/construction
• **Identifying prominent/trusted/expert actors**
• Identifying Spammers
• Discovering communities in social networks
• Link prediction
• Approximating large social networks
• Evolution of social networks
Yahoo! Answers
Question Life Cycle

Resolved Question

Why do all nuclear power plants use fission reactions and not fusion?

This was a weird question I was asked today. Does anyone know?
2 months ago

Best Answer - Chosen by Asker

Everyone above is correct. We can control fission, but we can't control fusion. We control fission by keeping dampers into the pile of nuclear material (e.g., U235). This interrupts the production of neutrons from each atom that splits. With fewer neutrons allowed to go on to split other atoms, the entire chain reaction is reduced to a controllable level.

Bottom line in fusion, we simply do not know how to interfere with the merging of two lighter elements (e.g., hydrogen) into one heavier one (e.g., helium). So when we reach the temperatures and pressures where fusion begins, it goes all or nothing. And the all is the blast of an H-bomb's fusion bomb. And, to date, no one knows how to contain an H-bomb to collect its energy on a continual basis.

3 months ago

Asker's Rating: *****

Thank you so much. This explained it perfectly.

Other Answers

Because nobody knows how to build fusion reactors.

3 months ago

Tom P

We can control fission reactions by damping them down with carbon rods.

3 months ago

dick<200

Controlled Fusion, converting hydrogen to helium, is still a dream. To date every effort to control a fusion reaction has failed.

3 months ago
Yahoo! Answers

Example of interactions between askers and best answerers

How to estimate the authority degree for each user?
PageRank?

Example: The category of “Programming”

- User B answers user A’s questions, which are about Java;
- User C answers B’s questions, which are about PHP;

➢ Is it possible to state that C is more expert than B?

- No, because: B and C have different expertise.
Proposed Approach

• The authority score of each user is simply the number of best answer of each users normalized so their square sum to 1:

\[ \sum_{i=1}^{N} (y_i)^2 = 1 \]

• \( y_i \) provide a relative score of the authority of each user in each category.

➢ We are interested in all sets of \( U_i \) having large values of \( y_i \).
Authority Score

• Example: Category of “Engineering”
Authority Score

[Graph showing distribution of Authority Score]
Automatic Identification of Authorities

Input: A set $U = \{u_1, u_2, ..., u_N\}$ of users
Output: A set $E = \{e_1, e_2, ..., e_d\}$ of authoritative users

1. For a given category, estimate the authority scores of each user;
2. Normalize $y_i$, where $\sum_{i=1}^{N} (y_i)^2 = 1$;
3. Estimate the pdf of the authority scores with $m = 2$;
   3.1. Apply FCM as initialization of the EM algorithm;
   3.2. Apply EM to estimate the parameters of the mixture;
4. Use the results of the EM algorithm in order to derive a classification decision about the membership of $y_i$ in each component.
Experiments

We conduct experiments on datasets which represent users’ activities over one full year for six categories:

<table>
<thead>
<tr>
<th>Category</th>
<th>% users who ask only</th>
<th>% users who answer only</th>
<th>% users who ask and answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td>65%</td>
<td>31%</td>
<td>4%</td>
</tr>
<tr>
<td>Biology</td>
<td>60%</td>
<td>36%</td>
<td>4%</td>
</tr>
<tr>
<td>Programming</td>
<td>66%</td>
<td>29%</td>
<td>5%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>64%</td>
<td>31%</td>
<td>5%</td>
</tr>
<tr>
<td>Physics</td>
<td>60%</td>
<td>34%</td>
<td>6%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>63%</td>
<td>32%</td>
<td>5%</td>
</tr>
</tbody>
</table>
Authoritative Users

Puggy

- Add to my Contacts  Block User

Member since: November 22, 2006

45,478 points  Level 7  49%

Best answer

Mathematics

About me: Just a 28 year old Canadian guy who loves tutoring math. I might want to teach someday.

As much as I love helping people understand mathematics, please refrain from e-mailing me your question in private; it's not fair that I should give special treatment to some and not others.

While I cannot guarantee to always be correct, I can (almost) guarantee a step-by-step solution where my mistake can easily be traced. Spotting of this error alone separates those truly willing to learn from those merely wanting the answer to their homework problem.

Mathematica

- Add to my Contacts  Block User

Member since: January 10, 2007

40,609 points  Level 7  72%

Best answer

Mathematics

About me: I have tutored math since 1993. My specialty is high school subjects - specifically Algebra, Geometry and Trigonometry. I currently work for a large test-publishing company.
Quality of Content

- The identified authoritative users generate high-quality content in Yahoo! Answers.
- Askers are very selective in choosing the best answerers.
Identifying Authorities in Online Communities

Workflow of the proposed approach.

- Users
- Features extraction
- Mixture modeling
  - Model the set of users’ feature vectors as a mixture of $C$ multivariate beta components
- Identification
  - Identify users associated with the selected multivariate beta component
- Selection
  - Select the multivariate beta component that contains vectors with highest feature values
Multivariate Beta Mixture Model

\[
F(\tilde{X}_i | \alpha, \tilde{a}, \tilde{b}) = \sum_{c=1}^{C} \alpha_c \ F_c(\tilde{X}_i | \tilde{a}_c, \tilde{b}_c)
\]

\[
F_c(\tilde{X}_i | \tilde{a}_c, \tilde{b}_c) = \prod_{d=1}^{D} f(x_{id} | a_{cd}, b_{cd})
\]

\[
f(x_{id} | a_{cd}, b_{cd}) = \frac{\Gamma(a_{cd} + b_{cd})}{\Gamma(a_{cd})\Gamma(b_{cd})} x_{id}^{a_{cd}-1}(1 - x_{id})^{b_{cd}-1}
\]
**Algorithm**

**ALGORITHM 2:** Authoritative users identification procedure

<table>
<thead>
<tr>
<th>Input</th>
<th>A set $U = {U_1, \ldots, U_N}$ of $N$ users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>A set $A = {A_1, \ldots, A_K}$ of $K$ authoritative users</td>
</tr>
</tbody>
</table>

**begin**

- For a given online community, estimate a feature vector $\tilde{X}_i$ for each user;
- Normalize $\{\tilde{X}_i\}$, as discussed at the beginning of Section 3;
- Apply Algorithm 1 to cluster the users into $C$ multivariate beta components;
- Use the results of the EM algorithm to decide about the membership of $\tilde{X}_i$ in each component;
- Select the multivariate beta component that corresponds to the highest feature values;
- Identify authoritative users in $U$ associated with the set of $\tilde{X}_i$ that belong to the selected component and store them in $A$;
- Return $A$;

**end**
Twitter data – 2012 Quebec election

• The data set consists of tweets posted between August 18, 2012 and August 20, 2012 (three days overall during the electoral campaign, including Quebec’s political party leaders’ debate which took place on August 19, 2012).

• 904 users; 76 users (8.4% of the whole data set) among them were labeled as authoritative and 828 users were labelled as non-authoritative
Twitter data – 2012 Quebec election

• Features
  • The number of followers of a user, which indicates the size of the audience for that user
  • The Followers to Followees ratio (F-F ratio), that is, the number of a user’s followers and the number of other people that the user follows (followees).
  • The number of retweets, which measures the number of times an author’s tweets were retweeted by other users
  • The number of mentions, which is measured by the number of times a user was cited or had her tweet replied to.
Twitter data – 2012 Quebec election

Density curves of several 2D user features combinations over Quebec Election
Twitter data – 2012 Quebec election

<table>
<thead>
<tr>
<th>Input features</th>
<th>Accuracy</th>
<th>CD</th>
<th>FA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td># of followers and F-F ratio</td>
<td>95.2%</td>
<td>97.3%</td>
<td>4.9%</td>
<td>0.774</td>
</tr>
<tr>
<td># of followers and # of retweets</td>
<td>97.3%</td>
<td>98.6%</td>
<td>2.7%</td>
<td>0.862</td>
</tr>
<tr>
<td># of followers and # of mentions</td>
<td>98.5%</td>
<td>100%</td>
<td>1.5%</td>
<td>0.921</td>
</tr>
<tr>
<td># of retweets and # of mentions</td>
<td>97.3%</td>
<td>94.7%</td>
<td>2.4%</td>
<td>0.857</td>
</tr>
<tr>
<td># of followers, # of retweets and # of mentions</td>
<td>97.3%</td>
<td>97.3%</td>
<td>2.6%</td>
<td>0.860</td>
</tr>
<tr>
<td># of followers, F-F ratio and # of retweets</td>
<td>97.4%</td>
<td>98.6%</td>
<td>2.6%</td>
<td>0.867</td>
</tr>
<tr>
<td>F-F ratio, # of retweets and # of mentions</td>
<td>97.3%</td>
<td>98.6%</td>
<td>2.7%</td>
<td>0.862</td>
</tr>
<tr>
<td>All features</td>
<td>99.2%</td>
<td>97.3%</td>
<td>0.6%</td>
<td>0.954</td>
</tr>
</tbody>
</table>

Performance results over Quebec Election data.
## Twitter data – 2012 Quebec election

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>CD</th>
<th>FA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>99.2%</td>
<td>97.3%</td>
<td>0.6%</td>
<td>0.954</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>99.2%</td>
<td>98.7%</td>
<td>0.7%</td>
<td>0.955</td>
</tr>
<tr>
<td>Bagging</td>
<td>98.8%</td>
<td>94.7%</td>
<td>0.7%</td>
<td>0.935</td>
</tr>
<tr>
<td>Decorate</td>
<td>99.4%</td>
<td>97.4%</td>
<td>0.4%</td>
<td>0.967</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>98.8%</td>
<td>94.7%</td>
<td>0.7%</td>
<td>0.935</td>
</tr>
<tr>
<td>MultiBoostAB</td>
<td>98.6%</td>
<td>94.7%</td>
<td>1%</td>
<td>0.923</td>
</tr>
<tr>
<td>ADTree</td>
<td>99%</td>
<td>96.1%</td>
<td>0.7%</td>
<td>0.942</td>
</tr>
<tr>
<td>Random Forest</td>
<td>99.3%</td>
<td>98.7%</td>
<td>0.6%</td>
<td>0.962</td>
</tr>
<tr>
<td>RBF Network</td>
<td>97.7%</td>
<td>92.1%</td>
<td>1.7%</td>
<td>0.875</td>
</tr>
</tbody>
</table>

Accuracies of compared algorithms on Quebec Election data.
Stack Exchange data
Unix & Linux
Q&A for users of Linux, FreeBSD and other Un*x-like operating systems.

questions 50k
answers 81k
answered 83%
users 72k

“Change top's sorting back to CPU” – asked 6 hours ago

Game Development
Q&A for professional and independent game developers

questions 22k
answers 41k
answered 92%
users 41k

“Why is it bad to hard-code content?” – asked 19 hours ago

Visit Site
<table>
<thead>
<tr>
<th>Input features</th>
<th>Accuracy</th>
<th>CD</th>
<th>FA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td># of answers and # of best answers</td>
<td>97.6%</td>
<td>76.6%</td>
<td>0.0%</td>
<td>0.868</td>
</tr>
<tr>
<td># of best answers and Z-score</td>
<td>95.7%</td>
<td>58.2%</td>
<td>0.0%</td>
<td>0.736</td>
</tr>
<tr>
<td># of best answers and # of votes received</td>
<td>90.7%</td>
<td>89.3%</td>
<td>9.1%</td>
<td>0.664</td>
</tr>
<tr>
<td># of answers and Z-score</td>
<td>95.3%</td>
<td>100%</td>
<td>5.2%</td>
<td>0.814</td>
</tr>
<tr>
<td># of answers, # of best answers and Z-score</td>
<td>97.7%</td>
<td>78.6%</td>
<td>0.1%</td>
<td>0.875</td>
</tr>
<tr>
<td># of answers, # of best answers and # of votes received</td>
<td>92.9%</td>
<td>100%</td>
<td>7.9%</td>
<td>0.743</td>
</tr>
<tr>
<td># of best answers, # of votes received and Z-score</td>
<td>92.1%</td>
<td>100%</td>
<td>9.8%</td>
<td>0.700</td>
</tr>
<tr>
<td>All features</td>
<td>99.1%</td>
<td>97.0%</td>
<td>0.6%</td>
<td>0.956</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Input features</th>
<th>Accuracy</th>
<th>CD</th>
<th>FA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td># of answers and # of best answers</td>
<td>97.6%</td>
<td>100%</td>
<td>2.6%</td>
<td>0.890</td>
</tr>
<tr>
<td># of best answers and Z-score</td>
<td>92.7%</td>
<td>94.8%</td>
<td>7.5%</td>
<td>0.718</td>
</tr>
<tr>
<td># of best answers and # of votes received</td>
<td>93.6%</td>
<td>92.8%</td>
<td>6.3%</td>
<td>0.739</td>
</tr>
<tr>
<td># of answers and Z-score</td>
<td>89.1%</td>
<td>96.9%</td>
<td>11.7%</td>
<td>0.635</td>
</tr>
<tr>
<td># of answers, # of best answers and Z-score</td>
<td>98.4%</td>
<td>94.8%</td>
<td>1.2%</td>
<td>0.920</td>
</tr>
<tr>
<td># of answers, # of best answers and # of votes received</td>
<td>98.9%</td>
<td>93.9%</td>
<td>0.5%</td>
<td>0.943</td>
</tr>
<tr>
<td># of best answers, # of votes received and Z-score</td>
<td>96.6%</td>
<td>92.8%</td>
<td>2.9%</td>
<td>0.842</td>
</tr>
<tr>
<td>All features</td>
<td>99.2%</td>
<td>95.9%</td>
<td>0.4%</td>
<td>0.959</td>
</tr>
</tbody>
</table>

(b)

Performance results of the proposed approach over: (a) Game Development data, (b) Unix & Linux data.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>CD</th>
<th>FA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>99.1%</td>
<td>97%</td>
<td>0.6%</td>
<td>0.956</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>99.1%</td>
<td>96.1%</td>
<td>0.6%</td>
<td>0.956</td>
</tr>
<tr>
<td>Bagging</td>
<td>99%</td>
<td>96.1%</td>
<td>0.7%</td>
<td>0.952</td>
</tr>
<tr>
<td>Decorate</td>
<td>98.4%</td>
<td>92.2%</td>
<td>0.9%</td>
<td>0.922</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>99%</td>
<td>96.1%</td>
<td>0.7%</td>
<td>0.952</td>
</tr>
<tr>
<td>MultiBoostAB</td>
<td>98.9%</td>
<td>97%</td>
<td>0.9%</td>
<td>0.948</td>
</tr>
<tr>
<td>ADTree</td>
<td>99.1%</td>
<td>97%</td>
<td>0.6%</td>
<td>0.956</td>
</tr>
<tr>
<td>Random Forest</td>
<td>98.9%</td>
<td>96.1%</td>
<td>0.8%</td>
<td>0.947</td>
</tr>
<tr>
<td>RBF Network</td>
<td>98.3%</td>
<td>94.2%</td>
<td>1.2%</td>
<td>0.919</td>
</tr>
<tr>
<td>SVM</td>
<td>98.9%</td>
<td>95.1%</td>
<td>0.8%</td>
<td>0.942</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>CD</th>
<th>FA</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>98.9%</td>
<td>93.9%</td>
<td>0.5%</td>
<td>0.943</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>98.9%</td>
<td>96.1%</td>
<td>0.7%</td>
<td>0.951</td>
</tr>
<tr>
<td>Bagging</td>
<td>97.9%</td>
<td>92.2%</td>
<td>1.3%</td>
<td>0.904</td>
</tr>
<tr>
<td>Decorate</td>
<td>98.3%</td>
<td>96.1%</td>
<td>1.3%</td>
<td>0.925</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>98.7%</td>
<td>96.1%</td>
<td>0.9%</td>
<td>0.942</td>
</tr>
<tr>
<td>MultiBoostAB</td>
<td>97.7%</td>
<td>94.1%</td>
<td>1.8%</td>
<td>0.897</td>
</tr>
<tr>
<td>ADTree</td>
<td>98.7%</td>
<td>94.1%</td>
<td>0.7%</td>
<td>0.941</td>
</tr>
<tr>
<td>Random Forest</td>
<td>98.3%</td>
<td>96.1%</td>
<td>1.3%</td>
<td>0.925</td>
</tr>
<tr>
<td>RBF Network</td>
<td>98.1%</td>
<td>96.1%</td>
<td>1.6%</td>
<td>0.916</td>
</tr>
<tr>
<td>SVM</td>
<td>98.7%</td>
<td>96.1%</td>
<td>0.9%</td>
<td>0.942</td>
</tr>
</tbody>
</table>

(b)

Accuracies of compared algorithms on: (a) Game Development data, (b) Unix & Linux data.
SNA Techniques

Prominent problems

• Social network extraction/construction
• Identifying prominent/trusted/expert actors
• Identifying Spammers
• Discovering communities in social networks
• Link prediction
• Approximating large social networks
• Evolution of social networks
Community Structure in Social Network
Graph Clustering

![Graph Clustering Diagram]

The diagram represents a graph with nodes and edges, and a matrix that likely indicates connections or clustering groups within the graph. The matrix shows binary values, possibly indicating the presence or absence of edges or clusters.
Algorithms based on Czekanovski-Dice Distance

Distance between two nodes

\[ \text{dist}(N_1, N_2) = \frac{|(S_1 \cup S_2)| - |(S_1 \cap S_2)|}{|(S_1 \cup S_2)| + |(S_1 \cap S_2)|} \]

S₁: number of nodes connected to N₁ (including N₁)
S₂: number of nodes connected to N₂ (including N₂)

Small distance ➔ High similarity
Czekanovski-Dice Distance

• Exemple

• \( \text{dist}(N_1, N_2) = ? \)

\( S_1 = \{N_1, N_2, N_3\} \)
\( S_2 = \{N_2, N_1, N_3\} \)

\[
\text{dist}(N_1, N_2) = \frac{|(S_1 \cup S_2)| - |S_1 \cap S_2|}{|S_1 \cup S_2| + |S_1 \cap S_2|} = \frac{3 - 3}{3 + 3} = 0
\]

• \( \text{dist}(N_3, N_4) = ? \)

\( S_3 = \{N_3, N_1, N_2, N_4\} \)
\( S_4 = \{N_4, N_3, N_5, N_6\} \)

\[
\text{dist}(N_3, N_4) = \frac{|(S_3 \cup S_4)| - |S_3 \cap S_4|}{|S_3 \cup S_4| + |S_3 \cap S_4|} = \frac{6 - 2}{6 + 2} = 0.5
\]
Czekanovski-Dice Distance

(a) Graph

(b) Similarity Matrix

(c) Dendogramme

(d) Clustering
Application

The Santa Fe Institute collaboration network
Application

Enron email network
Discovering Knowledge-Sharing Communities in Question-Answering Forums
Knowledge-Sharing Community

1. A knowledge-sharing community is defined by a set of askers and authoritative users.

2. Within each community, askers exhibit more homogenous behavior in terms of their interactions with authoritative users than elsewhere.

3. Authoritative users may belong to more than one community.
Knowledge-Sharing Community

Existing graph-based community detection methods are not appropriate for our study.
Example

\[
\begin{align*}
a_1 : e_1, e_2 \\
a_2 : e_1, e_2 \\
a_3 : e_2, e_3 \\
a_4 : e_2, e_3 \\
a_5 : e_1, e_2, e_3 \\
a_6 : e_1, e_2, e_3
\end{align*}
\]
Example

Modeling users interactions as a graph
The GRACLUS Algorithm
Modeling Interactions Between Users

We use a transactional data model to represent the interactions between askers and authoritative users.

\[
\begin{align*}
T_1 &= \{ e_1, e_2 \} \\
T_2 &= \{ e_1, e_2, e_3 \} \\
T_3 &= \{ e_1, e_2, e_3 \} \\
T_4 &= \{ e_2, e_3 \}
\end{align*}
\]

\[
\begin{align*}
T_5 &= \{ e_3, e_4, e_5, e_6 \} \\
T_6 &= \{ e_3, e_4, e_5 \} \\
T_7 &= \{ e_3, e_4, e_5, e_6 \} \\
T_8 &= \{ e_4, e_5, e_6 \}
\end{align*}
\]

• The first community is defined by \( T_1, T_2, T_3 \) et \( T_4 \)

• The second community is defined by \( T_5, T_6, T_7 \) et \( T_8 \)
Boolean representation of the interaction between askers and authoritative users.
The TRANCLUS Algorithm

- \( A = \{a_1, a_2, ..., a_n\} \) a set of \( n \) askers
- \( E = \{e_1, e_2, ..., e_d\} \) a set of \( d \) authoritative users
- \( TD = \{T_1, T_2, ..., T_n\} \) a collection of \( n \) transactions that summarizes the interactions of all askers \( a_i \) with the identified authoritative users.
Problem Definition

Given the set $A$ of askers and the set $E$ of authoritative users,

• Construct the set $TD$.

• Partition $TD$ into a set of disjoint clusters

  $C = \{C_1, C_2, \ldots, C_{nc}\}$

  - The identified clusters represent the communities we want to discover.
Criterion Function

\[
CF(C) = \frac{1}{n^2} \sum_{s=1}^{nc} \left[ \frac{1}{n_s} \sum_{e \in C_s} \left( \left( \text{occ}(e, C_s) \right)^3 \times Z(e) \right) \right]
\]

\[
Z(e) = \left( n - \text{occ}(e, TD) + 1 \right)
\]
The TRANCLUS Scheme

Input: A set \( TD = \{T_1, T_2, \ldots, T_n\} \) of \( n \) transactions
Output: A partition \( C = \{C_1, C_2, \ldots, C_{nc}\} \) of \( nc \) clusters

begin
  for each item \( e \) in \( TD \) compute the component \( Z(e) = (n - \text{occ}(e, TD) + 1) \);
  // Initialization phase
  while not end of the dataset file \( TD \) do
    Read the next transaction \( <T_i, \text{unknown}> \);
    Assign \( T_i \) to an existing or new cluster \( C_l \) to maximize \( CF(C) \);
    Write \( <T_i, C_l> \) back to \( TD \);
  // Refinement phase
  while \( move == \text{true} \) do
    \( move = \text{false} \);
    while not end of the dataset file \( TD \) do
      Read the next transaction \( <T_i, C_l> \);
      move \( T_i \) to an existing or new cluster \( C_t \) to maximize \( CF(C) \);
      if \( C_l \neq C_t \) then
        Write \( <T_i, C_t> \) back to \( TD \);
        \( move = \text{true} \);
  end
Application to Yahoo! Answers

(a) Biology.

(b) Chemistry.

(c) Engineering.
Content Analysis

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>{PHP, Website, HTML, JavaScript, Ajax, Java}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 2</td>
<td>{C++, net, games, Windows, Java, Microsoft}</td>
</tr>
</tbody>
</table>

(a) Programming.

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>{electricity, circuit, transistor, capacitor, battery, resistor, signal, amplifier}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 2</td>
<td>{mechanic, engine, motor, design, piping, fluid, machine}</td>
</tr>
</tbody>
</table>

(b) Engineering.

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>{cell, dna, blood, human, chromosome, gene, virus}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 2</td>
<td>{animal, mitosis, meiosis, cell, bacteria, chromosome, genetic}</td>
</tr>
</tbody>
</table>

(c) Biology.

- The clustered askers tend to post questions on closed related topics
Emerging Application

Influence of Social Networks on Product Recommendations

- Understanding the impact of social networks on market behavior
- Improved recommendation systems