PAPER REVIEW

Presented By: Udochukwu Nweke

CS734: Introduction to Information Retrieval

Dr. Michael Nelson

December 14, 2017

Paper 1

Exploring the Community Structure of Newsgroups

Christian Borgs Jennifer Chayes ∗ Mohammad Mahdian † Amin Saberi ‡

Date: 2004

Paper 2

Automatic Scoring of Online Discussion Posts

Nayer Wanas **Motaz El-Saban** **Heba Ashour** **Waleed Ammar**

Cairo Microsoft Innovation Center
Smart Village, KM 28 Cairo/Alex Desert Rd.
AbouRawash, Giza, 12676, Egypt
(+202) 3536-3207
{nayerw,motazel,hebaa,i-waamma}@microsoft.com

Date: 2008

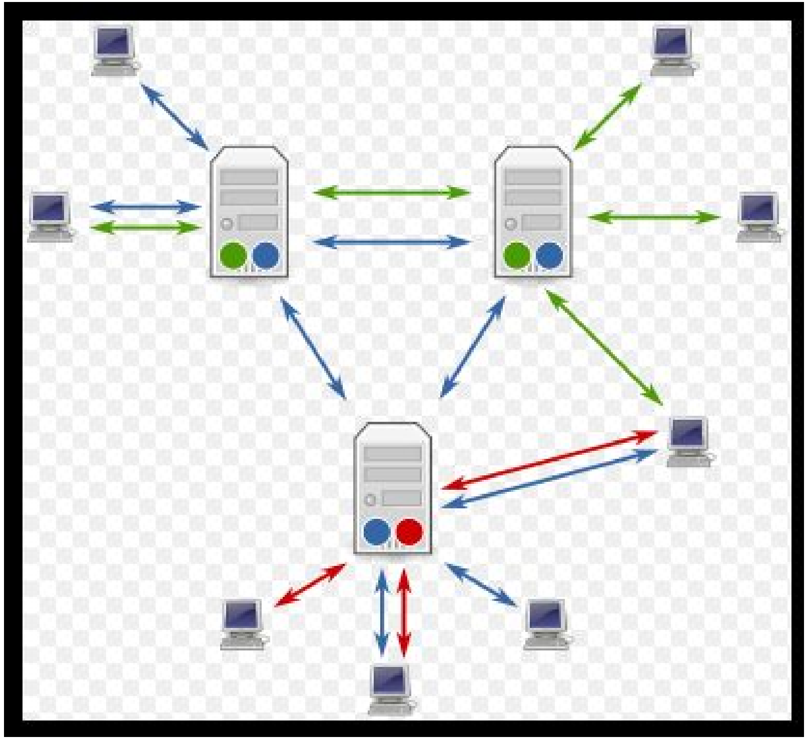
Objectives

- The recent interest in the structure of self-organized networks and various social network led to the study of Usenet
- Studied the cross-post structure of Usenet to organize and retrieve information stored in newsgroups

Usenet

- A worldwide distribution discussion system with over 50,000 newsgroups with different topics
- Users post messages or articles to different newsgroups
- The messages are distributed to other interconnected computer systems through different network
- The blue, green, and red dots on the servers shows the groups.
- The arrows between servers shows newsgroup group exchanges (feeds)
- The arrows between user and servers shows the group a user belongs to (post messages or articles)

Usenet servers and clients.



https://en.wikipedia.org/wiki/Usenet

Usenet: A Repository of Useful Information

- Usenet has become a big storage for information as a result of the rapid growth

- Because of this constant growth and the undefined structure, accessing newsgroup information has become difficult.

- Other attempts to explore usenet structure focused on the semantics of the contents. For example. Words in subject heading, group name, etc.

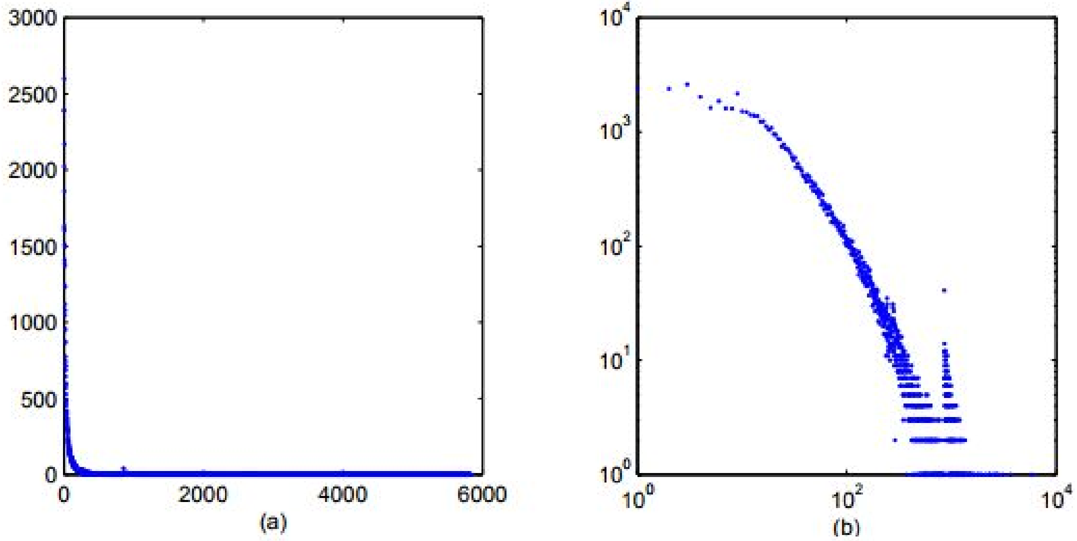
Cross-Post Graph Method

- Cross-post graph is a graph that shows an example of when messages are posted to two or more newsgroups at the same time.
- Cross-post graph shows a close relationship between a newsgroup and all other newsgroups
- Cross-post graph is close to a power law distribution. A relative change in one quantity causes a proportional relative change in the other quantity.

Cross-Post Graph

- The graph shows the number of cross-posts between a newsgroup and all other newsgroups

- That is, the probability that a newsgroup has x cross-posts with other newsgroups is proportional to $x^{-\alpha}$; here $\alpha \approx 1.3$



(a) (b)

Spectral Clustering Algorithm

- Spectral graph partitioning tool is used in many applications like web page partitioning.
- Cross-post graph G = (V, E) where V is the set of vertices corresponding to newsgroups and E is the set of edges corresponding to cross-posts
- G is a multigraph i.e. there may be several edges between two vertices
- The purpose of clustering is to partition the network into well-connected components

Graph Partitioning

- The purpose of clustering is to partition the network into well-connected components, the cut defined between two components is relatively sparse
- The commonly used measure for partitioning is : $\frac{\mathrm{cut}(S, \bar{S})}{\min(W(S), W(\bar{S}))}$
- Partitioning V into S and $\bar{S}$:
- Finding a cut that minimizes the above ratio is difficult

Heuristic Algorithm

- Heuristic algorithm allows to choose a splitting value s and divide the vertices into two sets based on whether or not the value assigned to them by v is greater than s.

- Best cut: Take s to be the value which gives the best cut according to the cut objective function in the equation .

$$\frac{\mathrm{cut}(S, \bar{S})}{\min(W(S), W(\bar{S}))}$$

- This method is recursively used until the size of each component is sufficiently small

Clustering Result

- The clustering percentage is 83.13%, while for a random clustering of the graph with the same distribution of cluster sizes this percentage is less than 1.53%.

- This comparison shows that the cross-post graph is indeed strongly clusterable, and the algorithm was successful

Paper 2

Automatic Scoring of Online Discussion Posts

Nayer Wanas Motaz El-Saban Heba Ashour Waleed Ammar

Cairo Microsoft Innovation Center
Smart Village, KM 28 Cairo/Alex Desert Rd.
AbouRawash, Giza, 12676, Egypt
(+202) 3536-3207
{nayerw,motazel,hebaa,i-waamma}@microsoft.com

Date: 2008

Online Communities

- Web applications that hold user-generated content
- User threaded discussions of sets of posts
- They form rich repository of collaborative knowledge
- Content is generated by user. Hence, the quality of post is determined by the user
- Finding useful information in this forum will be difficult and time consuming because of the rapid growth and unorganized structure
- Posts are manually rated by users and the scores help filter the forum content

Issues with Manual Assessment of Online Posts

- Reasonable portion of the threaded discussion could be overtaken by new threads before the value is identified and scored by a user
- Posts that come in after scoring are usually overlooked by moderators
- Quality of rating affected by the value initial posts
- Wrongly rated posts cannot be reversed

Automatic Assessment of Online Posts

- Various systems have attempted to automatically assess online post didn’t get a very good result because they assume that posts will follow linguistic rules and this is not true.

- Provided a better level for rating posts (low, medium, and high). This method is also conscious of linguistic phenomena pertaining to online discussion forums

- This is achieved by avoiding commitment to linguistic features and generating keywords from within the forum.

Post Scoring Metric

- The short and unorganized nature of online forum posts makes it difficult to evaluate.
- Another factor is the order and relationship with other posts.
- This work provides a seed value for each post through which a moderation process would rectify any misclassification
- In order to achieve this, a set of 22 features that are divided into five categories will be examined.

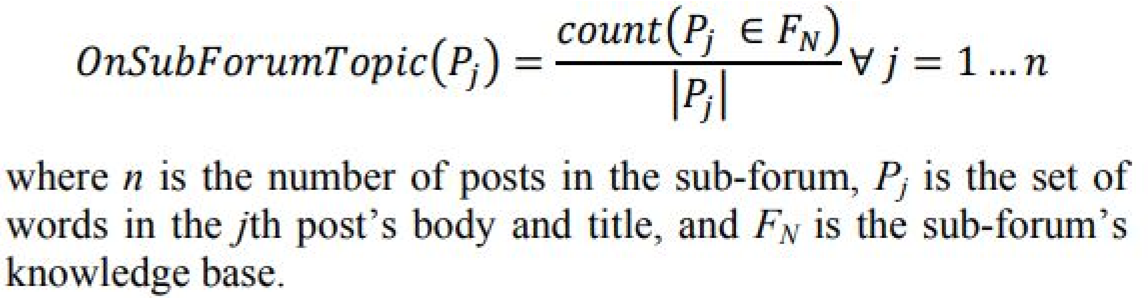
Generating Keywords

- These keywords are generated using a tf idf measure on a bag of words (BOW) combining all the words of posts in the sub-forum.
- These keywords are used to measure the important terms that distinguish the given sub-forum within online discussion forums
- The BOW of each post in the sub-forum (Pj) is then compared against the keywords description of the sub-forum

Relevance Feature

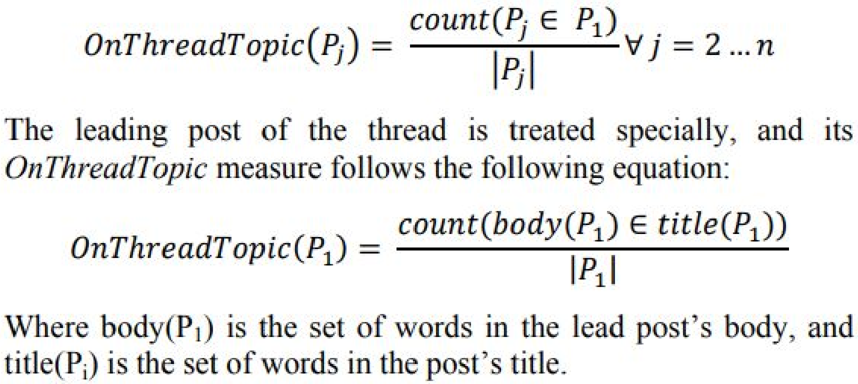
- Relevance is one of the most important feature that determines how a user perceives a post.
- Relevance determines the authenticity of a post to a thread and the subforum of the post.
- What determines relevance
- OnSubForumTopic: The degree a post remains relevant to the sub-forum
- OnThreadTopic: The degree of a post maintaining the relevance to the thread topic

Calculating OnsunForumTopic and OnThreadTopic



$$OnSubForumTopic(P_j) = \frac{count(P_j \in F_N)}{|P_j|} \forall\, j = 1 \dots n$$

where n is the number of posts in the sub-forum, P_j is the set of words in the jth post's body and title, and F_N is the sub-forum's knowledge base.



$$OnThreadTopic(P_j) = \frac{count(P_j \in\ P_1)}{|P_j|} \forall\, j = 2 \dots n$$

The leading post of the thread is treated specially, and its *OnThreadTopic* measure follows the following equation:

$$OnThreadTopic(P_1) = \frac{count(body(P_1) \in title(P_1))}{|P_1|}$$

Where body(P_1) is the set of words in the lead post's body, and title(P_i) is the set of words in the post's title.

Originality Feature

Originality goes hand in hand with relevance in deciding the value of a post
Two measures that determine originality are suggested:

- OverlapPrevious: This measures the degree of overlap between terms used in a post and the posts before in the same thread.
- OverlapDistance: This shows the separation, in terms of number of posts, between the previous post and that which has been judged as most overlapping by the OverlapPrevious measure.

Calculating OverlapPrevious and OverlapDistance

$$Overlap(P_i, P_j) = \frac{count(P_i \in P_j)}{|P_i|} \; \forall \, i > j, j = 1 \ldots n$$

Therefore, *OverlapPrevious(P_i)* is evaluated as

$$OverlapPrevious(P_i) = \max_{j}(Overlap(P_i, P_j))$$

The closer the overlapping posts are, the less value a post is

Forum-Specific Features

- The number of times a post is quoted and the number times a post is reproduced in common in an online forum. The features used to capture these aspects are: referencing and replies.
- The number of replies shows the value of a post.
- The value of a post increases if the text segment of the previous post is part of the current post.
- Quotation is a direction metric and will be evaluate with two features: CountBackwardReferences and BackwardReferences and CountForwordReferences and ForwardReferences

CountBackwardReferences and BackwardReferences

- CountBackwardReferences: This specifies number of quotation segments in the given post that are extracted from earlier posts.

- BackwardReferencing:This is used to quantify the value added to a given post by the quotations it contains.

$$BackwardReferencing(P_{ij}) = \sum_{i} \left(\frac{size\ of\ quoted\ text}{|P_i|} \times \frac{size\ of\ quoted\ text}{|P_j|} \right)$$

CountForwardReferences and ForwardReferences

- CounForwardReferences:This metric represents the number of times the post has been referenced in subsequent posts.

- ForwardReferencing: This feature aims to reflect the value added by a given post to subsequent posts that quoted it.

$$
\begin{aligned}
&ForwardReferencing(P_j) \\
&\qquad = \sum_{i} \left(\frac{size\ of\ quoted\ text}{|P_i|} \right. \\
&\qquad \times \left. \frac{size\ of\ quoted\ text}{|P_j|} \right)
\end{aligned}
$$

Surface Feature

The presentation of a post also determines it value because users will be attracted to pretty formatted posts and easy reading posts. The following metrics are used to determine surface feature:

- Timeliness: This determines how fast a user replies to a post
- $$Timeliness(P_j) = \frac{time\ difference\ between\ P_j\ and\ P_{j-1}}{Average\ inter - posting\ time\ in\ thread}$$
- Lengthiness: A posts that conforms to the maximum word count in a post is of a high value. The length is normalized by the mean length of posts in a given thread

$$Lengthiness(P_j) = \frac{|P_j|}{Average\ length\ of\ postings\ in\ thread}$$

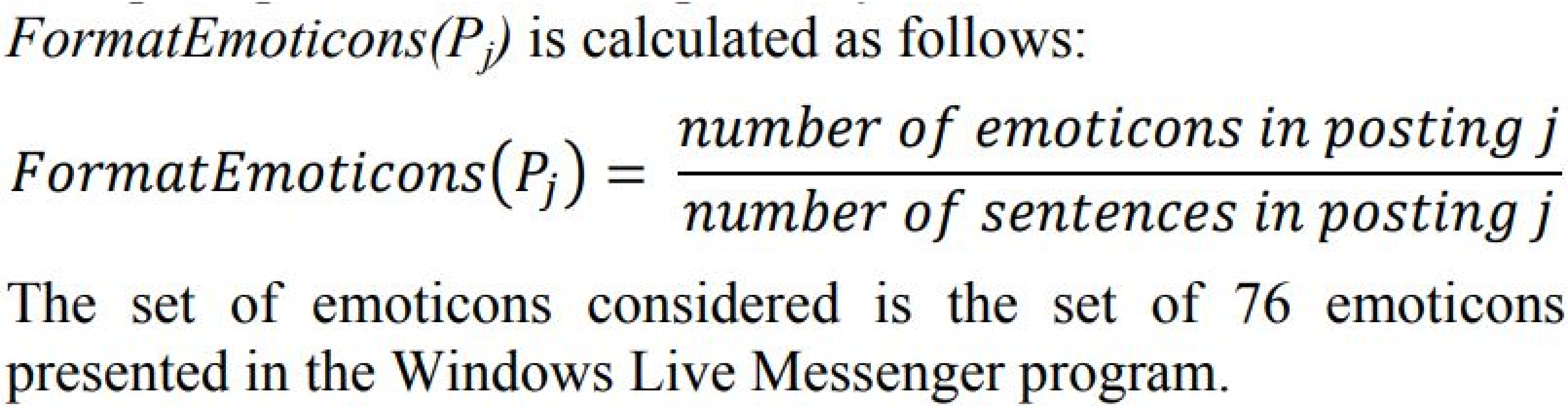
Surface Feature

- Formatting Quality: Too many formatting like emotions, capitalization, and punctuations undermines the value of a post. These three types are reflected using three features:
    - FormatPunctuation: Extensive use of creative punctuation affects users perception of a post. FormatPunctuation is calculated as follows:

$$FormatPunctuation(P_j) = \frac{number\ of\ chunks\ of\ consecutive\ punctuations\ in\ posting\ j}{number\ of\ sentences\ in\ posting\ j}$$

Surface Feature Contn’d

- FormatEmotions: Too much emotion in a given post conveys a level of emotion that affects the perception of the post by the user.



$FormatEmoticons(P_j)$ is calculated as follows:

$$FormatEmoticons(P_j) = \frac{number\ of\ emoticons\ in\ posting\ j}{number\ of\ sentences\ in\ posting\ j}$$

The set of emoticons considered is the set of 76 emoticons presented in the Windows Live Messenger program.

Surface Feature Contn’d

- FormatCapitals: Extensive use of capslock in a post conveys a tone that might affect its perceptions by users. For that reason, FormatCapital is calculated as follows:

$$FormatCapitals(P_j) = \frac{number\ of\ chunks\ of\ consecutive\ capitals\ in\ posting\ j}{number\ of\ sentences\ in\ posting\ j}$$

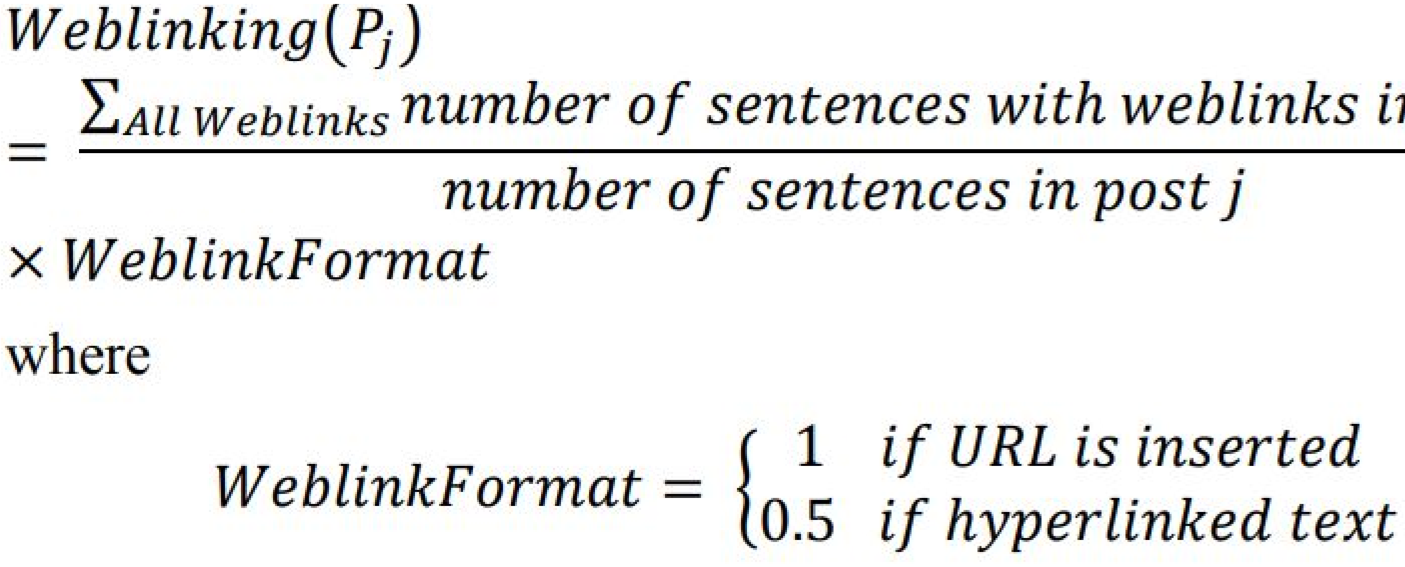
Posting Component Features

Since most dialogues on online discussion forums revolve around questions, web-links add value and credibility to posts. This forum elements are captured by two metrics: Weblinks, and Questioning.

- Weblinks: Including a web-link in a post adds value to a post and the value is composed of three factors:
  - Relevance of the web-link to the post
  - How the web-link is presented
  - The information about the web-link provided by the user
- These three factors comprise of two metrics used to determine the value of web-links present in a post namely Weblinking and WeblinkQuality

Posting Component Feature Contn’d

- Weblinking: This represents how well a user presented web-links in his posts. It is calculates as follows:



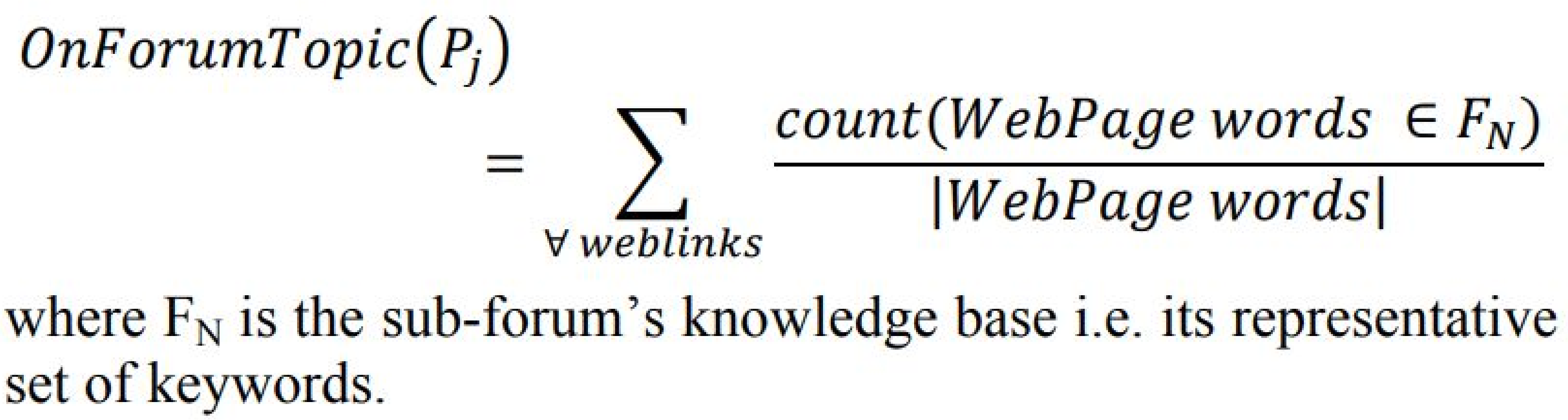
$$Weblinking(P_j) = \frac{\sum_{All\ Weblinks} number\ of\ sentences\ with\ weblinks\ in\ post\ j}{number\ of\ sentences\ in\ post\ j} \times WeblinkFormat$$

where

$$WeblinkFormat = \begin{cases} 1 & if\ URL\ is\ inserted \\ 0.5 & if\ hyperlinked\ text \end{cases}$$

Posting Component Feature Contn’d

- WeblinkQuality: This measures the similarity between the content of the weblink and the content of the sub-forum post. This is measured as follows:



$$OnForumTopic(P_j) = \sum_{\forall\, weblinks} \frac{count(WebPage\ words\ \in F_N)}{|WebPage\ words|}$$

where F_N is the sub-forum’s knowledge base i.e. its representative set of keywords.

Posting Component Feature Contn’d

- Questioning: Questions and their answers are one of the major features in online discussion forum.
- The value of a post is determined by the number of questions in the post
- The questions are assessed based on a template:
    - Question mark
    - Wh-questions

Slashdot Dataset

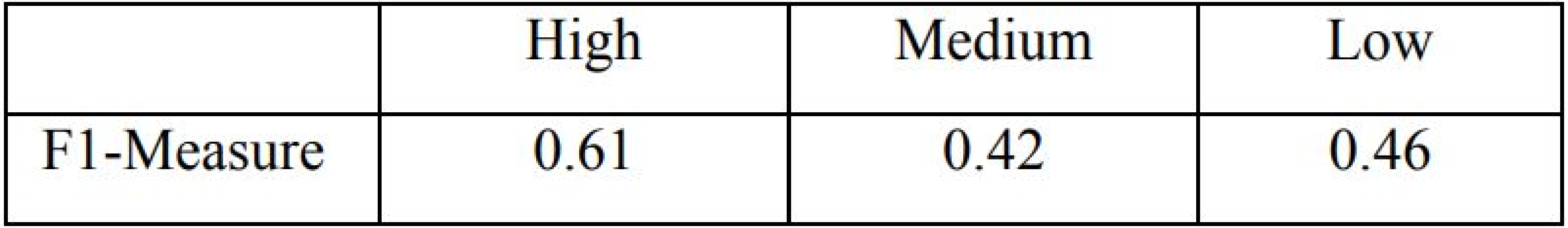
- Slashdot online discussion forum dataset was used for this experiment.
- 200 threads with a maximum of 200 posts each were selected from 14 sub-forums on slashdot.
- A total of 120,000 posts were scraped from the discussion forum
- Posts on slashdot are rated on a scale of -1 for irrelevant posts and 5 for relevant posts. The default rating for a non registered user is 0 and 1 for a registered user
- The final dataset is 20,008 rated posts which were clustered into three groups namely, low, medium, and high

Training Classifier

- A non-linear Support Vector Machine (SVM) classifier was trained using LibSVM[2], using RBF kernel to test the effectiveness of the features used.
- Five-fold cross validation was used on balanced data to evaluate the classifier performance. The performance was evaluated based on the accuracy and F1-measure
- The overall average accuracy and F1-measure of applying the classifier on the data set was 49.5% and 48.9% respectively.
- This level of accuracy is accepted for seed value for posts.

Result of Classification

Table 1: F1-measure on the three rating levels High, Medium and Low



	High	Medium	Low
F1-Measure	0.61	0.42	0.46

- It was observed that the performance on the posts rated “High” was significantly better than those rated as “Medium” and “Low”

- This is as a result of incremental nature of the rating policy implemented in Slashdot

Classification Result with individual Features

Forum-specific features has the most significant contribution to the overall accuracy

The results also show that the relevance and posting-component metrics have the lowest performance.

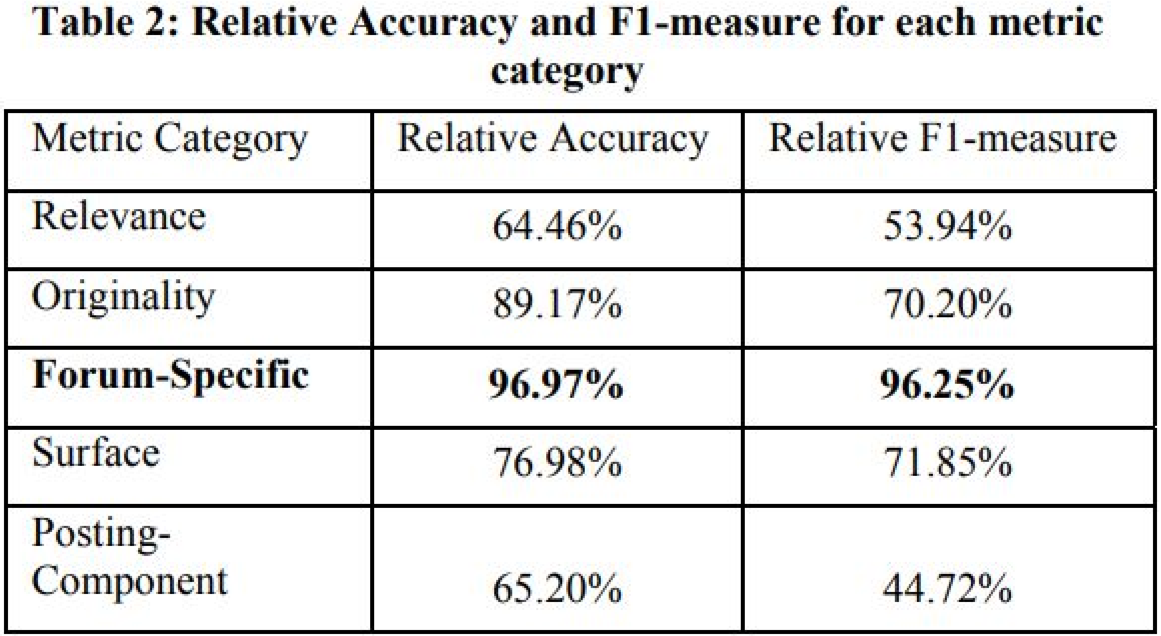


Table 2: Relative Accuracy and F1-measure for each metric category

Metric Category	Relative Accuracy	Relative F1-measure
Relevance	64.46%	53.94%
Originality	89.17%	70.20%
Forum-Specific	**96.97%**	**96.25%**
Surface	76.98%	71.85%
Posting-Component	65.20%	44.72%

Classification Result For All Pairs

- The combination of forumspecific and surface features was the best pair
- These two features score post at the face value and ignore the posting content.

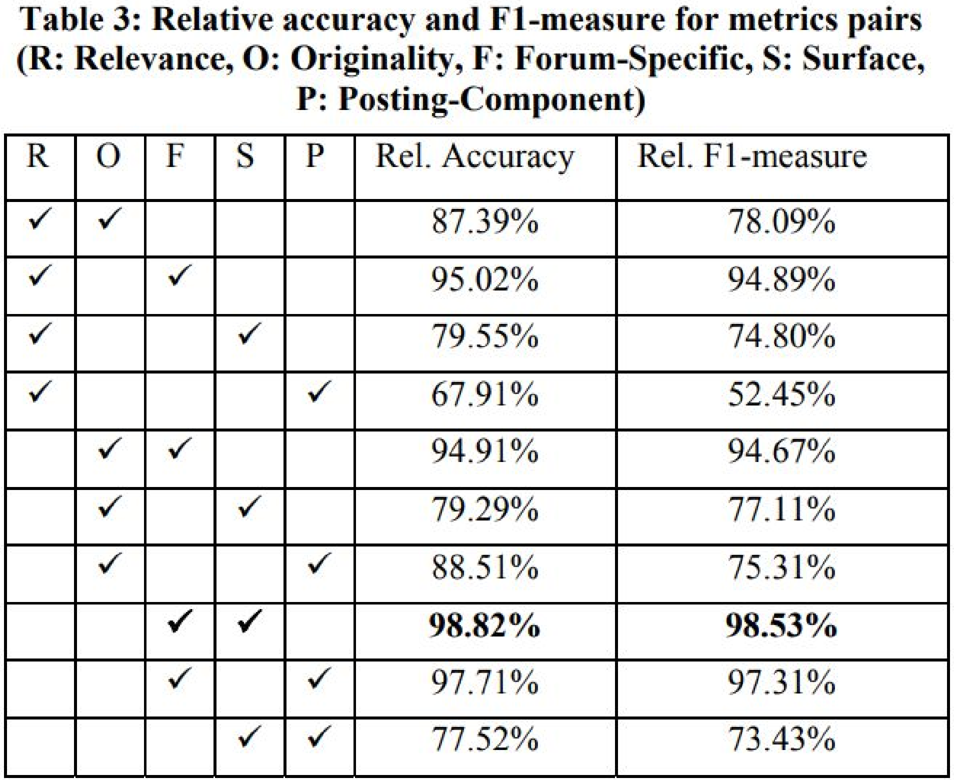


Table 3: Relative accuracy and F1-measure for metrics pairs (R: Relevance, O: Originality, F: Forum-Specific, S: Surface, P: Posting-Component)

R	O	F	S	P	Rel. Accuracy	Rel. F1-measure
✓	✓				87.39%	78.09%
✓		✓			95.02%	94.89%
✓			✓		79.55%	74.80%
✓				✓	67.91%	52.45%
	✓	✓			94.91%	94.67%
	✓		✓		79.29%	77.11%
	✓			✓	88.51%	75.31%
		✓	✓		**98.82%**	**98.53%**
		✓		✓	97.71%	97.31%
			✓	✓	77.52%	73.43%

Online Discussion Content Filtering

- Finding knowledge within an online discussion forum is a vital task. Relying on human rating through collaborative intelligence scheme to filter and organize post content is challenging. Many posts are not rated or are wrongly rated.

- Automatic rating is an alternative for manual rating

- In the course of evaluating relevance, a set of keywords were generated to describe each sub-forum

- These keywords can be used as browsing elements of posts in a given sub-forum

Relationship between Paper 1 and Paper 2

- Usenet, Newsgroups, Online discussion groups and other community generated repository have obtained recent focus concerning social networks and the structures of these repositories

- The rapid growth and structure of online discussion forum makes it difficult to access the information in the repository.