# Blog to Microblog

## Project Proposal CSCE-670

Vandana Bachani

February 16, 2012

**The Project**

The idea is to build a tool which can summarize articles (blogs, news articles) etc. based on the social context of the article. The ultimate goal is to summarize the content to a tweet (140 chars - pretty challenging), while an achievable target would be to summarize it in terms of important points (max 5) which capture the essence of the article from a social perspective.

The problem can be divided into 2 sub-problems, summarization of articles for which some social context exists (i.e. people have tweeted or commented, etc.) and summarization of new articles which do not have direct social context associated. The scope of this project is solving the first sub-problem and providing a decent for the second sub-problem.

**Motivation & Uses**

- The tool can be exploited as a new feature for twitter or any blog/newssite wherein user can create content and then generate a summary for the audience. It can improve user experience for twitter users (getting rid of user woes of having to express themselves in 140 characters) and encourage user engagement and content creation.
- Article summaries (if accurate) may be used by search engines as an extra link under the result to give the reader a gist of the content.
- Blog digests can be improved to include summaries for the followers.
- If we can apply similar principles to summarize social updates then I can see many apps like Siri taking advantage of the same.

**Related Work**

Summarization is a hard problem of Natural Language Processing domain. From an Information Retrieval perspective its about finding important phrases/sentences in the document which summarize the content well. There was a recent paper by Yang et.al[1] in SIGIR 2011 about Social Context Summarization, wherein they tried to find important messages in a document using a tweet-user-document graph based strategy. But they did not build any tool for the same and I don't think one exists which takes the social context into account. I look at the problem from a different perspective, i.e. of an implicit recommendation system, wherein the system needs to come up with a recommendation of important sentences in a document. This a new way of looking at the problem and also presents a novel application of recommender systems.

The challenges involve mapping the summarization domain problem to the recommendation domain problem with the correct set of parameters so that the approach gives meaningful results and, evaluating the approach as no standards exist for measuring the summary accuracy of such a system.

**Technical Details**

Given the social context of an article and the textual features, I plan to map the problem of summarization into a recommendation system problem.

- A document can be considered as a set of "sentences" which can map to "items" in the domain of recommendation systems.
- User tweets/comments act as ratings for the specific items of the document.
- The article can have some text context features such as zone weighting, weights based on similarity to the topic, uniqueness of items, considering only content-rich phrases, etc.
- A utility function is formulated based on the various features and the user social context (the ratings measure), which gives a utility to the sentences/phrases in the document and we apply the recommendation algorithms to get the recommendations of important phrases in the document which can be treated as the summary of the document from a social perspective.

I will be using the twitter dataset and extracting urls of documents that have been discussed on twitter (blogs, tech articles, news articles) etc. As twitter data is a sampled dataset and might not contain all the information, I will look at some tools like topsy.com (a social search engine) to get full social context of particular articles I might be interested in working with.

**Presentation and Evaluation**

The tool will have a graphical interface where it can be shown to work on documents by producing a summary (not necessarily grammatically coherent). If possible a 140 chars tweet as well. I will evaluate this using some measures mentioned in the paper[1] and also would present some results from manual user study or evaluation using volunteers to judge the accuracy of the system.

[1] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11). ACM, New York, NY, USA, 255-264. DOI=10.1145/2009916.2009954 http://doi.acm.org/10.1145/2009916.2009954