Enhancing Blog Clustering using Sentiment Analysis and Blogging Behaviour
Heba Ashour
March 5th, 2009
INTRODUCTION & BACKGROUND
Blog?

• An online journal that is frequently updated and intended for general public consumption.

• Blogs are an example of freely-expressed, unconstrained content.

• The nature of the blogging environment encourages readers to respond and comment on each others’ entries.
State of the Blogosphere*

• Over 106 million blogs exist & over 70 million are currently active

• Steady increase of 120,000 blogs per day on average

• Reading/writing blogs has become the 4th most popular activity on the web after email, search, and reading news

• 1,400,000 posts per day

Clustering?

- Dividing data into groups (clusters) that are meaningful and useful.

- It has been used extensively for information retrieval, e.g. for grouping search results of a query into meaningful categories.
A recent discipline at the crossroads of information retrieval and computational linguistics which is concerned with finding how subjective a certain document. This is accomplished by analyzing the amount of sentiment expressed in the document.

Blogs are opinionated in nature since they reflect the stance of their writer.
Blogging Behaviour

• The manner by which a blogger populates his blog.

• This is reflected by the following:
  – The topics he discusses
  – The frequency of his blogging on each topic
  – When did he discuss each topic?
  – How subjective is his writing?
  – Is his writing mostly expressing his own views or commenting on others?
PROBLEM & MOTIVATION
Problem & Motivation

• Information overload
• Arising need to aggregate blogs in a more meaningful way to enable better search, recommendation, summarization, and mining of their content.
OBJECTIVE
Objective

• Efficiently clustering blogs by using topical sentiment analysis, blogging times, and consistency as clustering features besides content.

• Achieving better results in the applications of search and blog recommendation than base-line content-clustering techniques
EXPERIMENTATION
ENVIRONMENTS
Exp Environment

• The experimentation will be carried out on 2 collections:
  - a collection of blog posts from the Arab world spanning 3 years from 1/1/07 – 1/1/09. They contain around 500,000 posts in 4,000 blogs in 4 languages. There’s also the set of comments on each post.
  - A collection of blog posts from Spinn3r.com spanning 2 months from 1/8/08 – 1/10/08. They contain 44 million posts in 6 million blogs in 40 languages. Their posts are also tagged with categories.
APPROACH
Approach

• Using a combination of tag-information and topical extraction techniques to detect topics discussed in a blog and their frequency.

• Detecting sentiment polarity with regards to every topic in the blog by analyzing sentiment at the partial sentence level.

• Clustering blogs based on the topics they discuss, when they discuss them, their consistency on discussing them, and the sentiments expressed about the topics.

• Comparing this method of clustering to base-line content-based clustering for search and recommendation.
Proof of Success

• The clusters generated will be first evaluated using cohesion and separation measures.
• These will be compared to the measures obtained from the base-line content clustering method.
• Also the two sets will be applied for search and recommendations and will be compared and evaluated based on precision and recall for humanly judged..
RELATED WORK
1
CLUSTERING BLOGS
Clustering Blogs

• Using Hierarchical Clustering:
  – Shepitsen et. al. – “Personalized Recommendation in Social Tagging Systems using Hierarchical Clustering”
  – Brooks and Montanez – “Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering”
2

SENTIMENT ANALYSIS IN BLOGS
Sentiment Analysis in Blogs

• Using Sentiment Orientation of words surrounding topic to give a sentiment score:
  – Kale et. al. – “Modelling Trust and Influence in the Blogosphere using Link Polarity”

• Using machine translation technology to perform sentiment analysis on english translation of foreign words
  – Bautin et. al. – International Sentiment Analysis for News and Blogs
Thank You

• Questions?