Presented at the DLSU Research Congress 2014
De La Salle University, Manila, Philippines
March 6-8, 2014

Synonym Based Tag Cloud Generation

Charibeth Cheng*, Therese Angustia, Mark Henry Ching,
Cara Andrea Cristobal, Germaine Marie Gabuyo
[1] *De La Salle University*
**chari.cheng@delasalle.ph*

Abstract: A tag cloud is a text-based visual representation of a set of tags which usually depicts the tag's importance in a given text. The presentation and layout of tags can be controlled so that features such as the size, font and color can be used to give some measure of importance of a given tag. Words that are used frequently will be displayed with an increased font size; while tags may appear in uniform or varying colors for aesthetics purposes or otherwise. The purpose of a tag cloud is to allow one to see, at a glance, the content of a document. Unfortunately, existing tag cloud generators produce clouds with tags that do not contribute in identifying the general content of a given document. These generators base the tags its frequency in the document. Thus, there may be tags, which are inflections of the same word, thereby populating the cloud with the same rootword. Furthermore, there may be frequently occurring non-stopwords, but are nevertheless non-discrimating (such as big, good, fine, etc.) SynTag is a tag cloud generator that uses syntactic and relational information of the words in the document in determing the relevant words/phrases in a document. With this, the generated tag cloud only shows tags which give focus on the content rather than illustrating the frequently occurring words. This system addresses the problems of existing term-based and frequency-based tag cloud generators.

Key Words: Tag Cloud, Synonym-based Tags, Word Relations

1. INTRODUCTION

A tag cloud is meant to visualize the words in a free form text. The relative importance of the words in the document are depicted through the size and color of the tags. Tags are identified based on the number of their occurrence in the document. Existing tag cloud generators include Dynacloud (Burkard, 2007), DrasticCloud[1], Wordle[2], and TagCrowd[3]. Figure 1.0 shows the tag cloud for the Wikipedia entry on Barack Obama[4].

Because tag cloud generators are term-based

[1] http://www.drasticdata.nl/DDDrasticCloud.php

[2] http://www.wordle.net/

[3] http://tagcrowd.com

[4] http://en.wikipedia.org/wiki/Presidency_of_Barack_Obama

and frequency-based, the following issues are encountered:

1. The same words with differing letter-case are considered as different tags. From Figure 1, we see 'President' and 'president' appearing as distinct tags.
2. The same words with differing plurality are considered as different terms. From Figure 1, we see 'American' and 'Americans'; as well as 'job' and 'jobs'.
3. Synonyms are not grouped together. We see this the tags 'jobs' and 'work'; 'problem' and 'crisis'. The importance of these concepts in the document is not properly because they are treated as unrelated words.

Non-discriminating words such as 'something', 'big', 'small', 'get', 'made' take up space in the tag cloud, because they occurred frequently in the document.

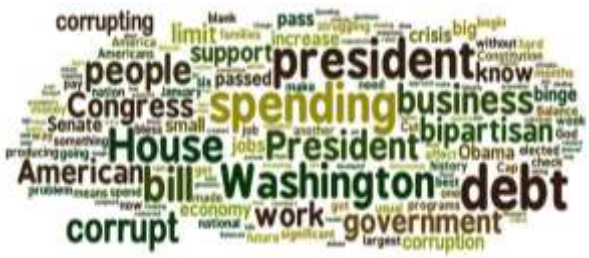

Figure 1. Example of Tag Cloud

To address these concerns, we built SynTag, which generates a tag cloud based on the syntactic and relations of the words in the document. The system addresses problems of existing tag cloud generators, which are term-based problem and frequency-based. By using term relevance and word similarity as basis, tags are identified. SynTag also uses color as a way to disambiguate homographs. Homographs are words that have the same spelling but have different meanings. Figure 2 shows a SynTag-generated tag-cloud.

Figure 2. SynTag Cloud

2. ARCHITECTURAL FLOW

Figure 3 shows the architecture of SynTag.

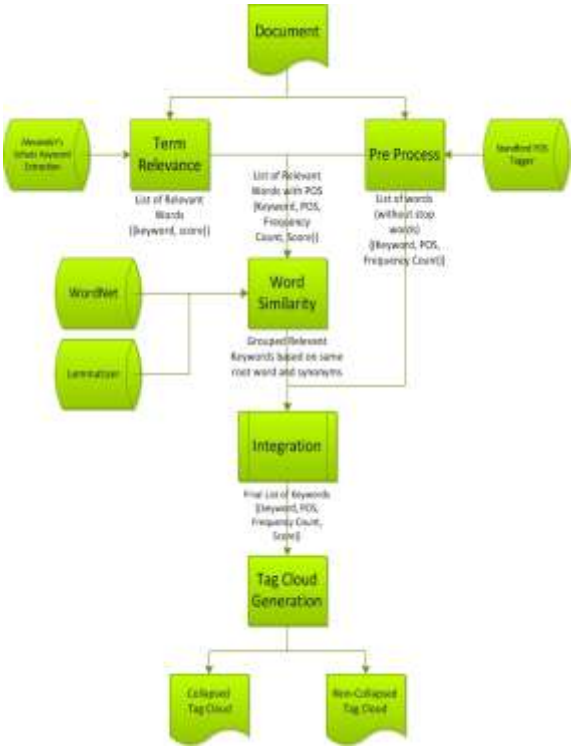

Figure 3. System Architecture

2.1 Pre-process Phase

During this stage, SynTag first removes all the stopwords in the document. The frequency of the

remaining words in the document are determined. Using the Stanford Part-Of-Speech Tagger (Toutanova et al., n.d.), each of these words are tagged with their part of speech. The output of this stage will be a list of words, with their corresponding part of speech, and their frequency count following the syntax <word>, <POS>, <freqcnt>.

2.2 Term Relevance Phase

In order to come up with the relevant words in the document, SynTag utilizes GATE (Schutz, 2008), or General Architecture Text for Engineering. The GATE provides each word with a score, indicating its relative importance in the document. The score given to each term will be ranging from 0 to 1, with 1 being most relevant. The result in this stage will be the relevant words, tagged based on what part of speech they belong, their frequency count, and their score in the form of <word>, <POS>, <score>.

2.3 Word Similarity Phase

This phase is run twice. During the first run, the list of relevant words produced from the Term Relevance phase is used. Related words from the list are identified and grouped together based on words having the same meaning and words that came from the same root word. Once these words are clustered, a candidate tag will be produced. Each cluster will have its own candidate tag.

To determine which tags are synonymous, SynTag uses RiTa WordNet[5]. On the other hand, words having the same root word were grouped using a lemmatizer, called ClearParser[6].

In the second run, the candidate tags identified in the first run are associated to the list of words generated during the Pre-Process phase, using the method described above. The output of this phase will be a Word Similarity List that includes the words, the corresponding part of speech, the frequency counts, the percentages, the synstes, the lemmas, and their matching groups following the syntax of <word>, <POS>, <freqcnt>, <percntge>, synset, lemma, group.

[5] http://www.rednoise.org/rita/wordnet/documentation/

[6] http://code.google.com/p/clearparser/

2.4 Integration Phase

Once the candidate tags are selected, the system now performs the Integration phase. In this stage, list generated by the Term Relevance phase and Pre-Process phase will be joint to come up with a list of clustered words, each having the same meaning and coming from the same root word, with their candidate keys.

To populate the clusters, candidate tags are compared to the list of words generated during the Pre-Process phase. Once a candidate tag matches a word from the list of the Pre-Process phase (both words have the same meaning or share the same root word), the matching word will then be included in the corresponding cluster of the candidate key it matches.

2.5 Tag Cloud Generation Phase

The last phase the system goes through is the Tag Cloud Generation phase, where the tag cloud for the document will finally be generated. Given the sets of words produced from the previous phases, this phase now assigns the corresponding sizes and colors for the words. The size will dictate the importance or frequency of a term, while the color, on the other hand, will represent its corresponding part of speech. Once done, words are now ready to be displayed in the tag cloud.

3. OBSERVATIONS

Consider below the tag clouds generated from an essay entitled "The Political Economy of the Philippines Under Marcos" by Kenji Kushida. Figure 4 shows the cloud generated by SynTag, while Figure 5 shows the cloud generated by TagCrowd.

SynTag categorizes words based on their root and synonyms, to form meaningful groupings of tags. Related words are positioned close together in the tag cloud. As seen the Figure 4, the following related word groups are positioned close to each other:

- power, ability, powers, force
- government, governments, governance, regime, politics
- commitment, commitments

- Philippines, Philippine

Figure 4. Tag cloud generated by SynTag

Furthermore, prominent phrases were also extracted from the document using POS and word relations. The following are some of the noun phrases identified as tags of the document:

- Local elite
- Marcos era
- Solid commitment
- Martial law
- Asset holders
- Credible threat
- External funding
- Long-term self-interest

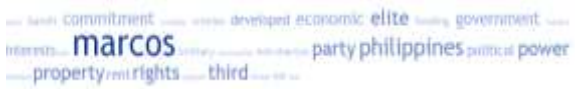

Figure 5. Tag cloud generated by TagCrowd

Figure 6. Tag cloud generated by Wordle

To differentiate the cloud generated by SynTag from other clouds, Figure 5 shows the cloud generated by TagCrowd using the same document, while Figure 6 is generated by Wordle.

The generation of relevant words is executed in the term relevance module. By utilizing the General Architecture for Text Engineering, or GATE (Schultz, 2008), we were able to compute the score of each word in the document. These scores describe the bearing of words, which is necessary for the system to use as the basis in determining the relevant words in the document. Furthermore, the output of this tool yielded to a conclusion that term frequency is not enough to be the basis of relevance of words in the document. This implies that the term relevance module is capable of determining relevant words even if they do not appear frequently in the document. Thus, performing this task allows the system to generate tags focusing more on term relevance than term frequency.

The grouping of words is based on word relationship; which is either via synonymy, or root words. To identify this relationship, we used (1) WordNet (Fellbaum, 1998), which determines synonymous words, and (2) ClearParser[7], which determines words that have the same root. This module performs a significant part in the system as it satisfies the main purpose of the study, that is to treat related words as one and prevent redundant words from appearing in the tag cloud. Figure 7 highlights the grouping of (power, ability, powers and force) and (governments, regime, government, governance and politics).

[7] https://code.google.com/p/clearparser/

Figure 7. Grouping of Related Words

Words having the same spelling but differs in meaning is a major concern particularly when such words appear in the tag cloud. This is addressed by the system through the application of colors where each part of speech is represented with a unique color in the tag cloud. Figure 8 shows how the different variations of the word Philippines/Philippine is represented in the tag-cloud.

Figure 8. Use of Color in the Tag Cloud

4. FURTHER WORK

Improvements particularly on the aesthetics of the tag cloud is needed. We propose adjustments to be made on the cloud's shape and size, as well as enhancements on the font and color. (Bielenberg and Zacher, 2006) presented a Tag-Cloud with circular form, where font size and distance to the center represent the importance of a tag, but where distance between tags doesn't represent their similarity. The center was chosen since it is defined for arbitrary polygons and it yields consistent layouts over a wide range of shapes. Therefore, we propose the generation of circular layout of tags to future system developments.

We also recommend a way to generate a tag cloud cloud that is dependent on user preferences, where the selection of fonts, colors and shape of the tag-cloud can be indicated. This will allow users to design the appearance of their tag clouds.

5. REFERENCES

Bielenberg, K and Zacher, M. (2006). Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation. Master Thesis submitted to the Program of Digital Media, Universität Bremen.

Burkard, J. (2006, August). DynaCloud – a dynamic javascript tag/keyword cloud with jquery. Available from http://johannburkard.de/blog/programming/javascript/-a-dynamic-javascript-tag-keyword-cloud-with-jquery.html

Burkard, J. (2007, August). Dynacloud v2 (more dynamic tag clouds) plus automatic related content. Retrieved Jan 29, 2014 from http://johannburkard.de/blog/programming/javascript/-a-dynamic-javascript-tag-keyword-cloud-with-jquery.html

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts, London, England: The MIT Press.

Gunn, N. (2010, April). Keyword tag cloud generator. Retrieved Jan 29, 2014 from http://www.seomoz.org/ugc/keyword-tag-cloud-generator

Herren, J. (n.d.). Tagcloud is getting an overhaul. Retrieved Jan 29, 2014 from http://tagcloud.com/

Marinchev, I. (2006). Practical Semantic Web – Tagging and Tag Clouds. *Journal Cybermetics and Information Technologies,* 33-39.

Schutz, A. T. (2008). *Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods.* Ireland, Galway.

Toutanova, K., Klein, D., Manning, C., Morgan, W., Rafferty, A., Galley, M., et al. (n.d.). *Stanford Log-linear Part-Of-Speech Tagger*. Retrieved Jan 29, 2014 from http://nlp.stanford.edu/software/tagger.shtml