Leveraging Collective Wisdom in A MultiLabeled Blog Categorization Environment

by

Magdiel F. Galan

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2015 by the
Graduate Supervisory Committee:

Huan Liu, Chair
Hasan Davulcu
Jieping Ye
Baoxin Li

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

One of the most remarkable outcomes resulting from the evolution of the web into Web 2.0, has been the propelling of blogging into a widely adopted and globally accepted phenomenon. While the unprecedented growth of the Blogosphere has added diversity and enriched the media, it has also added complexity. To cope with the relentless expansion, many enthusiastic bloggers have embarked on voluntarily writing, tagging, labeling, and cataloguing their posts in hopes of reaching the widest possible audience. Unbeknown to them, this reaching-for-others process triggers the generation of a new kind of collective wisdom, a result of shared collaboration, and the exchange of ideas, purpose, and objectives, through the formation of associations, links, and relations. Mastering an understanding of the Blogosphere can greatly help facilitate the needs of the ever growing number of these users, as well as producers, service providers, and advertisers into facilitation of the categorization and navigation of this vast environment. This work explores a novel method to leverage the collective wisdom from the infused label space for blog search and discovery. The work demonstrates that the wisdom space can provide a most unique and desirable framework to which to discover the highly sought after background information that could aid in the building of classifiers. This work incorporates this insight into the construction of a better clustering of blogs which boosts the performance of classifiers for identifying more relevant labels for blogs, and offers a mechanism that can be incorporated into replacing spurious labels and mislabels in a multi-labeled space.

DEDICATION

To my dear wife, who stood by my side from beginning to end. Without your love and support, this would not had been possible. To my kids..., shall they always follow their dreams. This was for you. A mis padres, y a toda mi familia..., gracias por todo su apoyo. Les quiero mucho.

ACKNOWLEDGEMENTS

It has been with great pleasure and forever be grateful for the assistance and support throughout this quest to the faculty, staff, and students at the Arizona State University. A lifetime experience that will forever be cherished.

I would particularly like to thank Arizona State University Professors Dr. Baoxin Li, Dr. Hasan Davulcu, Dr. Jieping Ye and Dr. Yi Chen, for their encouragement and support throughout this years. But most of all, to Dr. Huan Liu, whom without his guidance, faith, patience, and support, this great achievement would not had been possible.

I would also like to thank all the members of Data Mining and Machine Learning (DMML) group at Arizona State University. Each and every member were instrumental, in more ways than one, into this success. Thank You!

Finally, I would also have special thanks to DMML's ASU graduate Dr. Nitin Agarwal, whose initial collaboration paved the outcome of this work.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Blogging has evolved into a widely accepted and globally adopted worldwide phenomenon. A remarkable achievement considering blogging had only been introduced into the World Wide Web not less than a little over a decade ago. With only a handful of blogs, in the domain of a handful savvy programmers and few members of the academia on those early years, circa late 1990's, the popularity of blogging was catapulted by the advent of Web 2.0 (O'Reilly 2005).

Blogging had its humble beginnings with a human tone, while documenting personal experiences, events, accounts and the sharing of knowledge through a web page in the internet others could access and follow. On the early days of blogging, the blogs were simply web pages in which a selected few, with the capability and know how, would publish, addressed topics of interest and logged events, typically of personal interest in nature. These web-based logs were initially termed as *weblogs*, until eventually it was shortened to the term all too familiar as simply *blogs*.

These initial blogs lacked popularity, were not very user friendly, much less interactive. Maintaining the blog was rather cumbersome, as participants could only interact through emails, or following on threads, or message boards, typically controlled by a moderator. The process often required the thread trail to usually be manually handled, updated, unless the blogger possessed the technical know-how.

However, thanks to the improvements brought on by Web 2.0 changed the way in which users could interact and contribute to a blog. It transferred the domain from the hands of an almost exclusive technical savvy realm, to the regular masses.

The movement incorporated tools that facilitated the generation, maintenance and support of web styles pages configured as blogs, where it facilitated the chronological ordering of previous threads or postings, the linking and reference to other pages or blogs, the attachment of media sources and overall interaction. As blog popularity and use surmounted, others took notice, sprouting blog-posting services that assisted, facilitated, providing design, maintenance and archival of the postings. These services promoted, even enticed with opportunities for remuneration through advertisement.

Practically, the innovation brought by Web 2.0 made possible to anyone with access to a computer, with the ability to publish a blog. As many more joined the fray, communities of bloggers started to build, pursuing common passion and interests, and building in the process a highly interconnected and interdependent connected collective that is now recognized and termed as the *blogosphere*.

Since Web 2.0, blogs have further evolved, matured, not only limited to personal accounts, but also appeal to many other venues that have entered this space (corporates, franchises, special interest groups, etc.), thus propelling blogging into the juggernaut that is still today. Some have proclaimed the demise of the blog, Yet, hundreds of new blogs are started every day. Someone willing to share, or some group with a new product or services. The opportunity of remuneration. The thrill to be heard, be of purpose and significance. Old blogs die, new ones take their place. Is as dynamic as the evolution of people and products. Blogging has proven resilient, clearly stating it is here to stay.

1.1 Blogging Into Wisdom

The features and capabilities now embedded into Web 2.0 opened the door to regular users, fomenting what has been humanities long life quest. Since the dawn of man, human prowess can be traced to advances in military, defense, medicine and health, and technology, which for all intend and purpose, demonstrates an inherent need to communicate. From cave paintings to papyrus and tablets. From messenger pigeons to smoke signals and the pony express. From telegraphs and telephone, to satellites and cellphones. The human race is long driven to communicate, long to connect.

Perhaps driven by the sense to belong, people must communicate. To keep in contact with family and friends. To share opinions or views. To promote or advertise. To plan or coordinate. To educate or learn. To inform or entertain. To critique or praise. To influence or persuade. Thus communicate near and far, across the hall, and out of this world.

It is then to no surprise that with the advent of Web 2.0, the internet world was more than ready to adopt, eager to embrace, the new medium. This in turn spearheaded innovation, participation, and collaboration, enticing new players to flourish in this arena. These newcomers, spawned a surge of online media content in the form of wikis, and social sites such as for networking, like Facebook, bookmarking, like del.icio.us, online photo sharing, like Flickr, micro-blogging as Twitter, and many other such collaborating sites and services.

Users can now be defined by their digital footprint: the sites they visit, the apps users load, their likes and dislikes, the tags they generate, the tweets they tweet, the blogs they post. These various forms have become so ingrained, that have turned into a natural extension of who they are.They are the Internet, and the Internet is them.

Such is the path paved brought upon by the adoption of blogging. As blogging started to mature, it permeated society, it became an essential part of daily jargon and tech life. Such has been the extent, that in 2004, "blog" was declared "Top Word" of the year by Merriam-Webster. Journalism "by the people" or *citizen journalism* began to flourish, and in 2006 Time named "You" as Person of the Year, contributed by the growth of user-generated-content.

In following this new found passion, people not only had a new capability of generating content, but also had the opportunity, even encouraged, to enrich the existing media, whether textually, whether with media of their own, or both. A popular example is Google's image labeler, which is a social game that involves human subjects labeling images and retaining the most commonly used tags for a particular image. Another popular form of image tagging is geotagging, where tags of this kind represent forms of localization such as GPS coordinates, a town or city by name, or a point of interest. Other examples include tagging blog sites, or posts with relevant or descriptive labels, providing the necessary elements that would allow for search, cataloging, and alike.

As the blogging communities grew in following, so did the sharing and the exchange of opinions and ideas. Many very strong and passionate, many others divisive and combative. The blogs, with their colorful postings, were to become a truer reflection of the pulse of the people.

This was reflected in the concentrated discussions, the sharing of knowledge, and diversity of opinions on focused topics by the myriad of contributors. Inadvertently, this very process of "connectedness" alone by itself has resulted in the evolution of a new form of *collective wisdom*.

Collective wisdom is here defined as the shared, common knowledge arrived at, derived from, by a group of individuals. Some of the properties imposed to meet this criteria were outlined by (Suroweicki 2004). Mainly of all, it can be confidently ascertained people are wiser when their *collective* is individually poised, diverse and independent, cohesive and with the capability to be aggregated.

In retrospect, collective wisdom is not a new concept. Some credit its origins to as early as to the days of Aristoteles, who wisely first stated *"The whole is greater than the sum of its parts..."* As such, it has been the source of study in sociology, psychology, philosophy, economics, political science and various other venues since. With the implication that "many heads are better than one", it invited the notion that more can be achieved, ascertained, through a collective effort, than by any single individual, and in which an average, consensus could be achieved when taken as a whole. Accepting this stance, helps curb the noise from spurious thoughts when not in the norm in a task at hand.

Appropriately, that adage of *"everything that is old, is new again"*, is very fitting, as the concept has come in vogue, where the success for (Surowiecki 2005)'s New York Times best seller is credited with the resurgence of the term. This is evidenced by the philosophy very much adopted in corporate and business enterprises, where contributions achieved through collaboration, and diversity of capabilities and skills is highly sought after. Very fitting thanks to global economy, and a shrinking world

thanks to the internet, when commonality of thought has breached cultures and is without borders.

The case for the applicability of collectiveness into this work's assessment interpretation, is very much ratified by (Landemore and Elster 2012), which acknowledges that the collective wisdom concept is still as prevalent today, as it was during Aristotle's era, though with some noted enhancements, distinctions, in the order of scope/size of the audience or participants, the collaboration and interaction, but most affirming of all, the conceptualization of networks and connectedness, even offering conjectures into the contribution by social networks. The author ponders on the ideas and thoughts that are not monopolized by a mere few, but that the power and legitimacy is in the numbers, while still maintaining the essence of diversity, completeness and aggregation of judgement. Overall, laying the foundation, and providing validity into this work's collective wisdom assessment.

It should be noted that collective wisdom, is sometimes referred as group wisdom, wisdom of crowds, open source intelligence and co-intelligence. Though often used interchangeably, there are some subtle differences. Intelligence may entail specificity, be focused, and temporal, whereas group and crowd may suggest a distinction as a numerical body count. For this work's purpose, the properties in which envisioning wisdom is as transcendental, profound and unbounded, whereas collective would imply to be distributed and decentralized.

This work also incorporates (Hong and Page 2012), impossing some elements of sophistication and diversity as a condition that allows for it to be cohesive that could be aggregated as one, which seems to align with (Andler 2012)'s *thickly collective* as opposed to his designation of *thinly collective* where the aggregations are done to an individual assessment or conclusion rather than a collective one. This also align with

(List 2012) and (Davis-Stober et al. 2014) in the need to ensure that the collective is highly coherent, unbiased and not very off-centered, in which diversity is emphasized, where a great penalty would have to be paid if the elements in the collective are highly disruptive or misaligned to the corpus. Tolerating such discrepancies or follies, as (Briskin and Erickson 2009) refers to these instances, would introduce noise and variability, negatively impacting the collective wisdom's effectivity and outcome.

Collective wisdom has been credited with unusual feats, capable of obtaining the best possible approximation to a perfect solution, create associations, establish concepts and generate decisions. Its powers of intuition has been tested again and again. Sample of some popular deeds attributed to collective knowledge described by (Surowiecki 2005), (Landemore and Elster 2012), (Andler 2012), (Briskin and Erickson 2009) are the claims of correctly guessing the number of marbles in a jar, when guessing the weight of an ox at a fair. More actual, fresh example is the case for when a millionaire seeking contestant asks the audience for the correct answer on a popular game show, which most often than not, proven correct at a greater than 90 percent of the time.

What makes these examples most remarkable, the concept so tantalizing, and valid on its own terms, is the realization that these feats are without orchestration. They are the result of individual, independent contributions, which makes it most extraordinary, and in line with the properties for wise so described above. For each of these cases, the properties of diversity, cohesiveness, not adversarial, purposeful, and focused are present among others.

1.2 Needle in the Haystack - A Case for a Wiser Blog Clustering

As blogs increased in popularity and use, so did the increase in the variety of topics and subjects addressed. As bloggers matured, the blog topics evolved into more specific venues. Hence, blogs were no longer relegated into a collection of personal accounts and experiences. It was also now peppered with news media accounts, political venues, corporate sponsorship, and special interest groups. In a media, where everyone is given a voice, almost everyone is no short of opinion, whether in-depth discussions of current events happening throughout every corner of the world, political stance, world economics, pop culture, or more transcendental in nature.

The explosive increase in the number of blogs is not without consequences. A venue in which to communicate and share, to praise or critic, to inform, the potential for advertisement revenue, and the opportunity to "influence" products, services, public opinion and world stage affairs, appeals to many. Those seeking to join blogs they can follow and align with their passion and interests, may find it difficult to navigate the gamut of options, or worst, find none that meets their interests. Finding a best fit may prove more challenging to both users and advertisers. The heavy traffic, and growth, commands a need for automatic organization of blog sites that could facilitate navigate and explore this space.

One such organizational alternative is "clustering". It is a promising way to discover the composition of the Blogosphere and achieve the dynamic content organization being sought. Blog site clustering will not only helps to better organize the information, but will also facilitate convenient accessibility to the content. The clustering of blog sites will help optimize the search engines by reducing the search space. As such, clustering will allow to only search the relevant clusters and need not the entire

Blogosphere. A valued proposition for users, seeking a match to their passion and interests, as well as for advertisers, ensuring that their products and services are best aligned with the content will help them maximize the return on investment. That will help them maximize the best exposure and return on their investment.

Fortunately, a prominent feature of the Social Web is one in which many enthusiastic bloggers voluntarily write, or otherwise are perhaps encouraged, to catalog and label their posts with tag identifiers. The source for these tags could be various. The most common are of the free-form kind, in which users freely consider and choose keywords to represent to their post. In other instances, users may incorporate tags that will help associate their post to an intended audience, and in other times, select from a collection of suggested tags corresponding to a hierarchical type structure.

Tags aim to represent the essence of the blog post. Bloggers associate tags to their postings with the expectation to reach the widest possible audience, to draw attention and appeal to those who will share and value their thoughts and appreciate their ideas. Shared knowledge can then be drawn from these postings, when taking into consideration their labeling and level of similarity in content.

This work is to leverage on the blog's tags to exploit the properties from collective wisdom into the clustering of blog sites (Agarwal et al. 2008). In essence, it is to rely on the collective wisdom that can be achieved, when a collection of related, similar postings, from several independent bloggers, are associated to the same set of tags. More effectively, a particular collection of tags could then become defining of a post content.

Tags also play a significant role where they can lead for the blog to be discovered by advertisers, thus the potential for advertisement revenue. A key consideration to take into account, is that the Blogosphere follows a scale-free model and obeys

Long Tail distribution ((Anderson 2006)). *The Long Tail* is in reference to products and services that reside at the shallow end of the spectrum, away from the popular mainstream to which the majority consumers flock into on account of popularity, referred to as *The Head*. The reference to *The Long Tail* is to address those products and services residing away from *The Head*, as niche products that may cater to a specific interest of a very special few, which may offer unique, uncommon, customized and/or specialized products and services.

A vast majority of bloggers reside in the Long Tail and cannot be well targeted for otherwise potential business opportunities (i.e., niches) (Yin et al. 2012) (Shi 2013). To do better requires a good number of bloggers that can provide more data such as for targeted marketing traffic. This warrants a need for aggregating the Long Tail bloggers. Clustering various Long Tail bloggers to form a critical mass, will not only potentially expand a blogger's social network, but will also increase participation so as to move them from the Long Tail towards the Short Head. This could help the search engines to expand their result space and include results beyond just the Short Head. Including relevant clusters from the Long Tail in the search space would help in identifying those niches.

Clustering blog sites will invariably lead to also connecting the bloggers. Connecting the bloggers in the Long Tail may help in identifying *familiar strangers* (Agarwal et al. 2007). The underlying concept of familiar strangers is that they share some patterns and routines (or commonalities), although they are not directly connected. Clustering blog sites also helps promote the Web 2.0 new marketing 4Ps(Mootee 2001) *P*ersonalization, *P*articipation, *P*eer-to-peer, and *P*redictive modeling. Clustering is of the utmost importance as it will become quite critical into revealing and mastering

of the blogosphere's inner workings upon the Social Web that would serve the needs of users, producers, service providers, and advertisers.

Being that collective wisdom is at the core of this work, the intend is to maximize on the principles and properties and conditions stated that satisfy the collective wisdom earlier outlined such as diversity, focus and cohesiveness. Through the clustering effort, this work tackles the focusness of the blogs in accord to the wisdom properties, by similarity of their contents, would achieve in the process the ultimate task of combining similar blogs under the same category.

1.3 Wisdom from Tags? Better Tags for Better Clustering

A most notable contribution of the Web 2.0 has been to incorporate the capability for users to freely attach key words, or *tags* to their postings. This was a feature that was quickly popularized in bookmarking sites, like del.icio.us, and online photo sharing sites, like Flickr. The incorporation of the tags brought forward the expectation that it facilitated the search of related content. Their fast adoption prompted many other venues to follow suit.

The source for the tag selections could be various. The most common are of the free-form kind, in which users freely consider and choose keywords to best represent their object. In some instances, users may incorporate tags that will help associate their post to an intended audience. In other cases, selection may be influenced by popular, related tags, perhaps from a tag-cloud. Other options may include tags selected from a collection of suggested tags corresponding to a hierarchical type structure for a corresponding site.

The mere act of tagging humanizes a post. It provides an opportunity for a personal connection bond between the user and its post. The sense of meaning and purpose added authentication and validated a sense of "ownership", "control" inexistent prior to Web 2.0. But principally, tagging facilitates searching, indexing, acting as beacons for elements residing on *The Long Tail* (Landemore and Elster 2012) (Trant 2009).

The incorporation of the free form tagging has been the subject of much praise, but also apprehension. Collaborative tagging has made possible the generation of folksonomies, that is, taxonomies inferred by the loosely associations defined by common users, or folks, but may be influenced by what is popular, the signs of the times, consistency, completeness and much interpretation. Tag clouds reflect trends, popularity, but lack structure, relation or correlation among the tags within the said cloud. These challenges provide opportunity and are the source of continuous research in developing ways for maximizing their utility for discovery and navigation (Halpin, Robu, and Shepherd 2007) (Trattner, Körner, and Helic 2011) (Helic et al. 2011).

Selecting the proper tag to associate to a posting may prove most critical into categorization along other similar postings. Unfortunately, evidence suggests the role of the tags may not be well understood among some in the blogging community. Failing to exercise due diligence in a fitting tag selection, misses the opportunity to maximize, and take advantage of the full potential of the tags . Inaccurate tagging, whether intentional or unintentional, pays little justice to the blogger's postings, their potential followers, and advertisers.

A properly crafted tag is much source of consideration. Choosing a fitting tag can be both daunting and very challenging, particularly when considering its potential effect with regards to search engines, and in the attracting of followers and advertisers.

Care should be taken to avoid categorizing with labels that may be too generic or may loosely represent their contents.

In this tag-based wisdom dependent clustering scenario, a major factor that could hinder the quality of the clusters would be the wisdom exercised by the blogger into the selection of the associated tags that would most accurately represent their posting. Relying on proper tagging would be a necessity to support the quest to achieve quality clusters. As such, the attention focuses into how to best address potential tagging inconsistencies.

Fortunately, the environment is not quite the free-form leading to folksonomies just discussed. For one, it is not a product of collaborative tagging, as this assertions are independent from other bloggers. Another factor in favor is that the tag space is hierarchically bounded. Nevertheless, it may still be considered as a slightly different variation of (Trant 2009) notion of *narrow* folksonomy, in the sense of consisting of a finite tag set collection, and the assignments are implemented unknowingly, independent and uninfluenced or coerced by others. As such, considering this alternative, takes into account that there could be instances of inconsistencies in tag labeling since after all, this is a human process, and considering the freedom to choose a tag, though narrow choice it may be, it may still be inconsistent with respect to others in a similar space, either by choice, interpretation, or by design, introducing as such, spurious labeling in the process.

Being that collective wisdom, is at the core of this work, the intend is to maximize on the principles and properties and conditions stated that satisfy the collective wisdom outlined at the start of the chapter such as diversity, focus and cohesiveness. This work introduces a framework rooted on collective wisdom. Leveraging on strong

tag relation helps minimize the effect brought on by spurious labeling that could still be present despite a narrow tag space.

This work builds an environment, in which the tags are interpreted as labels in a feature space constructed from the blog contents into classifiers. The inherent label structure within the collective wisdom achieved from linking subsets of related tags as they conform to the blogger's labeled postings into building a better classifier. The outcome of this process will result in better aggregating *The Long Tail*, while enriching and diversifying the collective wisdom properties.

Chapter 2

RELATED WORK

2.1 Blog Clustering

The surge in blogs population brought forth by the emergence of Web 2.0 naturally drew the attention of researchers to the blogosphere, sparking interest onto the bloggers and their postings. This work focuses on research aimed at clustering of the blogs. The serach uncovers that authors in (Tseng, Tatemura, and Wu 2005), (Qamra, Tseng, and Chang 2006), (Chin and Chignell 2006) and (Lin et al. 2006) explore ways to identify communities. Such blog community based clustering rely on user induced connections in the underlying blog network to identify communities. However the resulting blog clusters only identified the community structure of the blogosphere and may not necessarily helped in clustering blogs of similar contents. A closer attempt to contents is done by (C. Brooks and N. Montanez 2006) utilizing the blog tags for hierarchical clustering.

Though content or topic based clustering of web documents and text has been widely studied, content based clustering of blogs has not been studied widely. Work by (Bansal et al. 2007) performs semantic analysis in order to discover topic trends, with the goal of identifying clusters that persists over time. The clusters are based on identifying bi-connected components in a graph. In (Xu and Zhang 2004) and (O. and Etzioni 1998) web document clustering has been done based on the K-Means algorithm (MacQueen 1967).

Apart from K-Means, agglomerative and hierarchical clustering has also been used for document clustering. Such is the case for (Yin, Han, and Yu 2006), which uses a hierarchical structure for linkage based clustering measured as by similarities of other objects linked to a pair of objects, where objects can refer to authors, papers, links, and web sites. Similar work is accomplished by (Xin Li 2006), where the authors use hierarchical clustering to try identify communities by establishing connections per the co-occurrence of words and entities in entities such as web pages and blogs. Authors in (Dubes and Jain 1988) present a review of the clustering algorithms and both (Cutting et al. 1992) and (Steinbach, Karypis, and Kumar 2000) have provided reviews of document clustering techniques.

The mentioned clustering algorithms can be directly applied to blogs by considering blogs as web documents. The adopted vector-space model can be used to encode the blogs represented as term frequency vectors for the similarity matrix using Singular Value Decomposition (SVD), to which apply a clustering technique such as some form of K-Means or hierarchical clustering and/or variations of them. However by doing this it would be ignoring the many unique characteristics of blog which would aid in obtaining a better clustering.

A significant consideration is that blog sites are not as rich in text and structure as professionally carefully crafted web documents. Most blog sites, or web "personal" logs as the name implies, are personal accounts, opinions, ideas, thoughts, and expressions that have less content and generally not well-authored. However, labels or tags assigned by humans (both bloggers as well as readers), also known as the collective wisdom, make them special and different from web documents. As such, traditional web documents keyword clustering algorithms mentioned above would fail to return

good results due to blog's sparsity and curse of dimensionality. Therefore novel techniques are required that leverages the enormous collective wisdom available.

2.2 Leveraging Tag Information

Collective wisdom as represented by the labels or tags provided by humans have been previously used for various tasks like search and retrieval, and recommender systems. The human annotation provided for web pages and blogs provide valuable metadata for use in search. Websites like 'del.icio.us', 'Flickr' and 'YouTube' use such user provided metadata in the form of collaborative tagging for search and retrieval. Since large amount of such metadata is available even in the blogosphere, it can be leveraged for search and retrieval operations. Authors in (Hotho et al. 2006) provide an algorithm to search using the tag information. In (Hayes, Avesani, and Veeramachaneni 2007) the authors have used the tag information for a blog recommendation algorithm.

While it has been proven the utility of the tags, it has also been highlighted the criticality of having tags that accurately depict and represent the blog's post content. Some of the proposals to facilitate some of the challenges in assigning labels to ensure accurate tagging representation are through various forms of tag recommenders. In (Mishne 2006) the author proposes AutoTags, which offers the bloggers a set of tags used by similar blogs, and weighted by tags previously used by the blogger. In (Sood and Hammond 2007) goes a step further with TagAssist, by pre-stemming the tags, edging AutoTags in their results. A variation of these is conducted by (Hart, Johnson, and Stent 2009), who achieves better results by limiting only to tags previously generated by same blogger. A potential shortcoming on these approaches

is that they rely on the consistency, and accuracy of the tags they base to make the recommendation in the first place. A different approach is evaluated by (C. H. Brooks and Nancy Montanez 2006) which compares the similarity of documents clustered using from popular tags to clustering through top keywords from the same documents per *Term Frequency - Inverse Document Frequency* (TF-IDF) criterion.

Though the use of 'collective wisdom', and the role of the tags have been studied as just mentioned, there is still opportunity for improvements in terms of using a greater variety of user generated data (like user provided labels in blogs), improving the quality of the clusters, through better label understanding and categorization, and for more kinds of applications in the blogosphere (like clustering of the blogs in the blogosphere) and the web in general.

2.3 Multi-Label Environment

As contents on the internet get more complex, diverse and rich, a single label can no longer satisfactorily do justice to the contents, particularly, for search engines scanning throughout the web. As such, many of the fundamental learning methods targeting a single class may not be applicable, without some type of variation. Hence, this type of understanding has prompted significant research as of recent. In (Zhang and Zhou 2013), and (Sorower 2010), a compilation of Multi-Label learning algorithms are provided.

This work identifies user driven links generating collective wisdom through the various label combinations, for as in a multiple label environment, where even though the individual labels may sometimes appear unrelated, are nevertheless paired in some form of seemingly unnatural existence. Items like poems-and-music, or cookies-and-

milk may seem natural fit, whereas others may require a little bit of a "stretch" of the imagination (i.e. animals-and-crochet). Work from (Zhang and Zhang 2010) and (Alaydie, Reddy, and Fotouhi 2012), explores such label dependency in a general context, whereas (Papadopoulos, Kompatsiaris, and Vakali 2010) incorporates a hybrid graph scheme for related tags in a folksonomy. Though similarity to the links this work build through collective wisdom, the technique incorporates means where it is minimize such, perhaps less natural, spuriously occurring label pairings.

The goal for this work is to re-categorize, or complement, a perhaps non-very descriptive label, with a more search-informative one. This will use blog posting text to learn which, from a collection of labels, would be a good more descriptive fit to replace with. Because of the complexity associated with a multi-label classification environment, works from (Bi and Kwok 2013) explores how to best perform this task efficiently. Fortunately, the task is simplified from the fact that the sampled blogs for the test environment were collected from a blog site directory that contains a hierarchically finite label pool (BlogCatalog[1]). BlogCatalog is a directory of blog sites that allows bloggers to record and label their blog sites under a given label hierarchy structure. Works from (Tsoumakas and Katakis 2007), and (Alaydie, Reddy, and Fotouhi 2012), make reference on techniques in support of hierarchical type labels.

[1]http://www.blogcatalog.com/

Chapter 3

BLOG CLUSTERING - A LABELED APPROACH

3.1 Introduction

This work seeks to achieve the clustering of blog sites. More formally, given a collection of m blog sites, $S_1, S_2, ..., S_m$, to construct k disjoint clusters of the m blog sites, such that $k \leq m$. Purposely exploiting the inherent collective wisdom while forming clusters of these blog sites. The collective wisdom is available in the form of predefined labels attributed to each blog site by the blogger.

When evaluating the label space for this environment, it is encountered that seldom does a blogger confine his blog to a single label that would entirely and uniquely describe its contents. It is more common to find blog sites which are frequently tagged under multiple labels. This multi-label environment imposes additional challenges that need addressing before embarking into this quest of forming clusters out of the labeled posts.

With the proposed label-based framework, requires exploring a way for clustering in which to demonstrate that the clusters thus obtained through the labeled approach are more meaningful and representative in their context as compared to traditional approaches, which continuously try overcome their inherent shortcomings like,

- Text clustering suffers from the curse of dimensionality and sparsity (Devaney and Ram 1997).
- The similarity measure does not capture the semantic similarity very well (Huang, Ng, and Jing 2006).

- The clusters thus obtained are sometimes not very meaningful (Huang, Ng, and Jing 2006).
- User needs to specify the number of clusters *a priori* which could be hard to anticipate (Song and Park 2006).

In order to leverage on collective wisdom for the clustering strategy, must tackle the challenges of the conventional clustering approaches described above. Since the label tags are the fundamental cornerstone of this work's approach, the quality of these clusters depends on the wisdom exercised by the bloggers into the selection of the associated tags that would most accurately represent their postings. This work addresses these concerns in the coming sections, where it will be the focus of attention into demonstrating a proof-of-concept technique that will allow relying on the associated posting tags, thus increasing the confidence in the quality of the generated clusters in the process.

3.2 Naïve Label Clustering

Clustering blogs by leveraging solely on their labeled categorization is a challenging proposal. This restriction dispenses from conventional clustering techniques. Under such constraint, an instinctively natural naïve approach would be to gravitate toward grouping together all blog sites that have been tagged under a specific label, or labels, as to belonging to the same one cluster.

In entertaining this option, it is at least best enforce a systematic approach. Therefore, the process is to follow a path in which to systematically first address clustering blogs categorized with a single label. This is then followed into clustering

those with two labels, and explore their relation to the previous iteration, and so forth.

This initial naïve step can be accomplished quickly and easily. Since typically most blog postings are tagged with greater than one label, the number of blogs with single label should most likely be quite small. This step will invariably yields to as many clusters as there are distinct single-labeled blogs. However, one consideration not taken into account on this approach is that there could be various single-labeled clusters that might be related, and should best be set together. Nevertheless, for the sake of this process, they are kept distinct.

The next step in this naïve approach would be to address the blogs with more than a single label, starting with those as label-pairs. But much to the contrary to the initial step, with its simpler blog to label-as-a-cluster assignment, this simple heuristic would not suffice as it is soon discovered that this approach faces with various options at this stage, any of which, or combination of them, could possibly be pursue. Case in point, consider a blog which is categorized with a pair of labels, both of which have been previously populated with correspondingly single labeled blogs. For this blog, it needs to be decided to which of the two distinct labels clusters should assign this new two-labeled blog. Furthermore, it is to consider the condition in which are identified several other blogs with same pair of labels. If significant enough blogs in this case, it may be perceived that each should be granted their own cluster-label status. Should consider revising the naïve approach, branding this blog set as their own category, could also declare the two previously independent label-clusters to now belong to this new cluster.

It should come to reason there would be additional sets of two-labeled blogs that would fit as their own cluster. Chances are, some of these two-labeled clusters may

share common labels, whether among themselves, or clusters prior generated. Should they be combined, many of these clusters may now be joined as one. By perpetuating the trend of joining clusters when sharing the same label, may very well end up with probably relatively very large clusters as the process progresses sequentially from two, to three-labeled blogs and greater, whose only relationship for many of the blogs contained within the cluster may simply be reduced to the sharing of a label, regardless of content.

Throughout the naïve process, this process has focused on matching blogs solely on their label categorization, purposely ignoring the blog's content, until eventually come to the realization this is not a sound strategy. As trying to scale, it becomes more apparent it only grows in difficulty to definitely decide to which of the many labels the blog can be individually assigned, and prevent at the end a scenario in which all blogs will eventually become part of a single large cluster.

Let instead explore a strategy which takes into account the blog's content, in order to establish similarity among the labels that should provide a relation among the blogs in the cluster. Clustering the similar labels can be better formulated as an optimization problem.

Assume to have t labels, $l_1, l_2, ..., l_t$ and are clustered into k clusters, $C_1, C_2, ..., C_k$, then optimal clustering is obtained if, for any two labels l_i and l_j,

$$\min \sum d(l_i, l_j), \forall (l_i, l_j) \in C_m, 1 \leq m \leq k, i \neq j \tag{3.1}$$

$$\max \sum d(l_i, l_j), \forall l_i \in C_m, \forall l_j \in C_n, 1 \leq (m, n) \leq k, m \neq n \tag{3.2}$$

Here $d(l_i, l_j)$ refers to a distance metric between the labels l_i and l_j. The first formulation minimizes the within-cluster distance between the cluster members while

the second formulation maximizes the between-cluster distance. Finding efficiently an optimal solution for the above min-max conditions is infeasible.

Existing work like (Xu et al. 2004) proposes a method for clustering based on maximum-margin hyperplanes through the data by posing the problem as a convex integer program. The hard clustering constraint is relaxed to a soft clustering formulation that can be feasibly solved with a semidefinite program. In a probabilistic approach, data is considered to be identically and independently drawn from a mixture model of several probability distributions (McLachlan and Basford 1988). An expectation-maximization (EM) based approach is used to first estimate conditional probabilities of a data point (x) given a cluster (C) by ($P(x|C)$) and then find an approximation to a mixture model given the cluster assignments. K-means is an approximation to EM based clustering approach. Another approach to cluster the blog sites is based on the tags assigned to the blog posts and the blog site. Each blog site can be profiled based on these accumulated tags. A simple cosine similarity distance metric could be used to find similarity between different blog sites. However, the vector-space model of the blog sites based on the tags is high-dimensional and sparse. The use of a singular-value-decomposition (SVD) based clustering algorithm as the baseline serves to avoid the curse of dimensionality.

In recalling the original intent of categorizing blogs based on their labels. It is by staying within these bounds, that in the process have completed reviewing two label-based models. Although both models have their own merits, in the process have also discovered they carry a few, though rather significant, shortcomings.

The enticing simplicity of the naïve ways, is overshadowed by its blind clustering drawback, under penalty of same label blog, where blogs are clustered regardless of whether or not they should belong together. Whereas when entertaining the blog's

content to help assess their togetherness, donned in return, is the complexity of having to address the limitations of a highly dimensional and very sparse environment, inherent of the label similarity technique's vector-space model.

3.3 Wisdom Based Clustering

The alluring concepts of "simplicity" and "similarity" of the naïve ways are intrinsic elements at the very core of the proposed alternative blog site clustering approach, which leverages on "collective wisdom". In this process, it is derived the knowledge that is inadvertently generated by the bloggers when labeling their postings. As it is often the case, bloggers typically specify more than one label for a particular blog site. In most instances, it would be very limiting to simply categorize a blog with a single label, as it would not be as descriptive, and certainly, constraints their efforts from possibly reaching their target audience and obscures search engines. Hence, as often, more than one label is penned to a blog.

This seemingly innocuous action has the unwitting effect of establishing a link between said labels, in a form that they are as consciously interpreted by the blogger, within context, for the corresponding posting. This brings an unique opportunity, in which to capitalize upon this interpretation and represent it into a *label relation graph* for the labels associated to a blog, a sample instance of which is depicted in figure 1. In this instance, predefined labels like `Computers and Technology`; `Computers and Internet`; `Computers and Blogging` were linked by the bloggers. The quantities associate to each of the links represent the link strength as the number of instances in which such pairing occurred for this example.

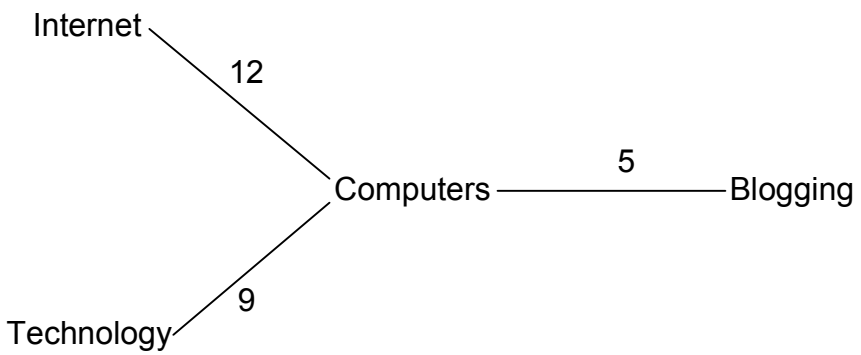

Figure 1. Instance of a Label Relation Graph.

In this process, it can also be observed that some bloggers may annotate their blogs using labels in hierarchical fashion. As can be observed from the sample label distribution illustrated in figure 2 of these cases, bloggers may often use `Personal` as a label descriptor, in combination with other labels for varied interests. For many of these cases, `Personal` represents an item that relates, or is on top, of their "personal" interest, and subsequently followed by more related, descriptive labels. Just as in `Personal`, and similar hierarchical-type instances, these type of labels may not be helpful in capturing the nuances of bloggers' intent and as a result, need to refine the label descriptor by identifying and aggregating the related labels. These type of cases, has been recognized, and normally referred as the *topic irregularity problem* where bloggers frequently use the same label descriptor to define their blog which in fact contains blog posts of varied interests. To overcome this case-conditions would require that different labels with similar themes be connected even when a blogger does not list his/her blog under all these labels.

The path for achieving collective wisdom is created when the process can identify multiple bloggers that unequivocally establish a link between same labels. The number of blog sites that create links between the various labels is termed as *link strength*,

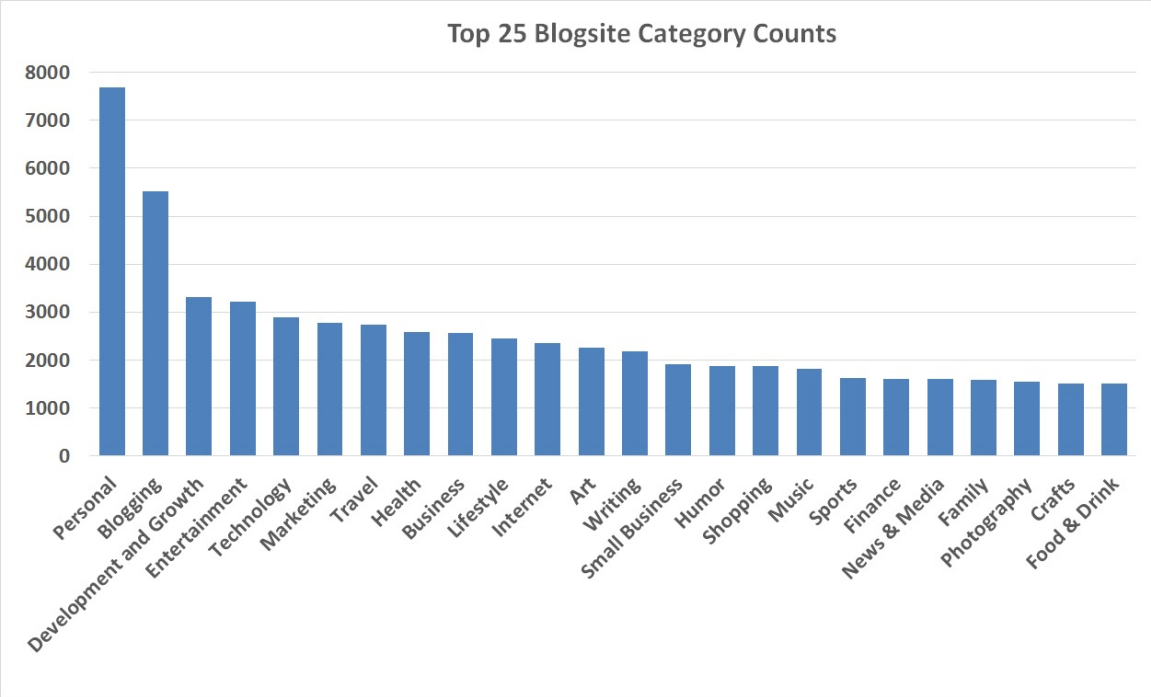

Figure 2. Distribution of blog sites with respect to top labels.

which could be treated as the edge weights of the label relation graph. Using this label relation graph, different labels can be clustered or merged. This collective wisdom based approach, will be referred as *WisColl*. Next is to experiment with different thresholds for the link strength in the experimental section. To visualize the label relation graph thus obtained, a visualization and analysis tool, Pajek[2] is utilized. Once the label relation graph is computed after thresholding, performing afterwards label clustering using k-means and hierarchical clustering algorithms and compare their results.

WisColl is time sensitive and adaptive to the current interests, since the labels of a blog site could change depending on what the blogger is blogging about. This results in dynamic, as well as adaptive, clustering. Every time new blog posts appear, either

[2]http://vlado.fmf.uni-lj.si/pub/networks/pajek/

there will be new edges appearing in the *label relation graph* and/or link strength changes, as blogger specifies different labels. As such, the clustering results would change, to reflect current state, keeping results fresh and relevant.

3.4 A Conventional Approach

As a clustering algorithm for the baseline, the process involves clustering the blogs using the blog post text and then find the predominant label for each cluster. The 'Vector-Space' model is used to encode the blogs with each blog being represented as a term frequency vector. Singular Value Decomposition (SVD) and the cosine similarity measure are then used to obtain the similarity matrix for clustering.

The vector-space representation for each blog is constructed to find the term frequencies in the blog posts of each blogger. For each blogger up to five blog posts are available and thus extracted and using these posts a 'blog-term' matrix is constructed. This work is restricted to only consider English postings. The following pre-processing steps are applied to the terms obtained by blog posts before constructing the matrix:

1. Trim white spaces and punctuation marks, token scrubbing is performed on the blog post text;
2. All the terms are *stemmed* using the Porter stemmer to obtain their morphological roots; and
3. The stop words are removed from the remaining list of terms.

After the preprocessing steps, using the resulting normalized terms the blog-term matrix B ($m \times t$ matrix, with m bloggers/blog sites and t terms) is constructed. Latent semantic analysis (Deerwester et al. 1990) is performed on this matrix to

obtain the lower dimensional semantic representation of each blog. This step required decomposing the blog-term matrix using SVD (Deerwester et al. 1990).

$$B = USV^T \tag{3.3}$$

The blogger-term vectors were then projected into the semantic feature space by selecting the top k singular values and the corresponding singular vectors from U and V. The reconstructed blog-term matrix is of rank k.

$$B_k = U_k S_k V_k^T \tag{3.4}$$

In the experiments, the best performance is achieved by selecting top 25 eigenvectors. In the resulting matrix B_k each row corresponds to one blog and is represented by the vector $d_i = (d_{i1}, d_{i2}, ..., d_{it}) 1 \leq i \leq m$. The $m \times m$ similarity matrix S was then constructed by finding 'cosine similarity' between term vectors corresponding to each pair of blogs. The $(i, j)^{th}$ element of S gives the similarity between blogs i and j and is given by,

$$S(i, j) = (d_i \times d_j)/(\|d_i\| \times \|d_j\|) \tag{3.5}$$

Once obtained, the similarity matrix clusters of bloggers/blogs can be easily visualized. Clustering is achieved by setting a threshold τ for similarity. A link between two nodes is considered weak if the similarity is less than τ. When the weak links are removed, clusters start to emerge. By identifying the predominant labels for the nodes in each cluster, the cluster labels can be identified.

Chapter 4

WISCOLL - LEVERAGING TAGS INTO CLUSTERING

4.1 Introduction

The collective wisdom tag based clustering approach will be referred as *WisColl*. To demonstrate the validity of WisColl, sample data collected from a blog site directory available at BlogCatalog[3]. This validation will serve as template to further test other blog and blog-like sources.

BlogCatalog is a directory of blog sites that allows bloggers to record and label their blog sites under a given hierarchical structure. The directory structure of BlogCatalog is a relatively shallow tree, with 33 nodes having no children. The maximum depth of the hierarchy is 3 and only two nodes have such depth. To test validity, a series of experiments with varying granularity of structural information. Bloggers submit the blog sites to BlogCatalog. Each site is authored by a blogger. Each blog site contains some blog posts of which snippets of the last 5 postings are displayed on the BlogCatalog.

Figure 3 illustrates the approach to data collection and experimentation. To test the technique, a drawing from a pool of bloggers is to be as the source. To this set of bloggers, it is applied the following clustering techniques: the baseline approach (refer to A Conventional Approach section), which incorporates singular value decomposition (SVD), and the approach to draw from wisdom collectivism, that identifies commonality among bloggers. This work uses Hierarchical (i.e. agglomerative), and K-Means

[3]http://www.blogcatalog.com/

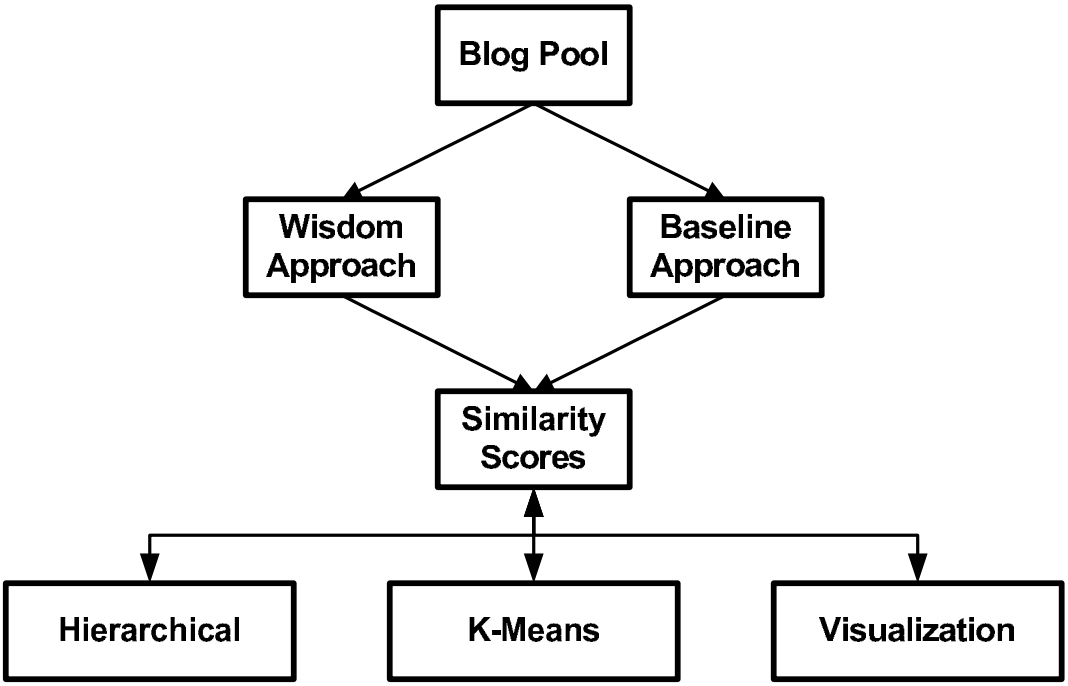

Figure 3. AnalysisTree.

clustering techniques to make possible comparison between the two approaches, and also the use Pajek's nodes graphing feature to help visually validate the analysis results. The intent throughout this chapter is to ensure and demonstrate that WisColl is a viable and effective clustering technique.

To collect the BlogCatalog data for this study, the process is initiated with 4 bloggers from different labels as the starting points and crawled their social networks, recursively, in a breadth-first fashion. These bloggers were selected to belong to the most popular labels (i.e., having largest number of blog sites) at BlogCatalog. For each blogger thus crawled (uniquely identified by their blogger IDs), the process collects their blog site's URL, title, and labels, the blog post tags, snippets, title, and permalink, and the blogger's social network information, i.e., their friends.

Along with evaluating collective wisdom, in the process, it is also evaluated the

structural properties of the labels, i.e., since the labels have a nested hierarchical structure, in order to help assess what level gives the best clustering results. To perform this, three different datasets are constructed:

1. *Top-level*: The hierarchical structure of the labels is known *a priori*. For this dataset, the labels are abstracted of all the blog sites to their parent level labels. For example, `Family` is a child of `Personal`. So all the blog sites that are labeled `Family` are relabeled as `Personal`, thus abstracting their labels to the parent level. Note that the maximum depth of this hierarchical structure of labels is 3 and in the process, abstract the labels to the highest parent level label. There are in total 56 labels after abstraction.
2. *All-label*: This variant of the dataset does not abstract the label information. It considers the full hierarchical structure of the labels. There are in total 110 labels at all the levels of the hierarchy.
3. *One node-split*: According to the distribution of blog sites in various top level labels, illustrated in Figure 4, `Personal` has the largest number of blog sites[4]. Hence, best split `Personal` into its child labels, to reduce the skewed distribution.

Note that the approach presented here is intended to work for any blog dataset with user specified metadata like labels or tags. As such, the outcome is designed to evaluate the following:

- What granularity of label hierarchical structural information generates best clustering? For this is necessary to study the clustering results for the three variants of the dataset indicated in the design method.

[4]For the sake of space constraint and the analysis presented here, best is to limit the labels, shown in this chart, that have at least 1000 blog sites.

- Which one of the clustering approaches: a) link-based clustering approach, (that leverages collective wisdom), or b) the baseline approach, performs best per K-Means, Hierarchical, or Visualization clustering?

Before delving into the parameter tuning of the various clustering methods issues, best study the effect of different link strength thresholds for WisColl. The aim is to fix the threshold for link strength for the rest of the experiments based on the results of this study. Thus, the *All-label* variant of the dataset for the threshold experiments is to be used.

4.2 Link Strength

The process is to experiment with different thresholds for the *All-label* link strengths[5] range of values. These values have been captured in Table 1. The table shows the range of values within the collection, and their distribution, for all of the 456 line pair values, that can be obtained from all of the 110 *All-label* node combinations.

Selecting a link-strength, or threshold, may result in a network re-structuring and reduction depending on the value selected. The re-structuring occurs as a result of removing those links whose line values are below the selected threshold, which can cause a cluster to transform into a smaller cluster and may possibly spawn additional clusters. The cluster size reduction occurs when removing any newly generated isolated nodes, or clusters, that were once linked to at least one other node in the cluster they once subscribed, but belong no more, as their connecting link was removed for falling under the threshold.

[5]number of blog sites that links same pair of labels

Table 1. *All – label* Link Strengths

Line Value	Frequency	Freq(pct.)	CumFreq	CumFreq (pct.)
1	320	70.1754	320	70.1754
2	62	13.5965	382	83.7719
3	33	7.2368	415	91.0088
4	15	3.2895	430	94.2982
5	7	1.5351	437	95.8333
6	7	1.5351	444	97.3684
7	6	1.3158	450	98.6842
8	0	0.0000	450	98.6842
9	2	0.4386	452	99.1228
10	1	0.2193	453	99.3421
11	3	0.6579	456	100.0000
Totals	456	100.000		

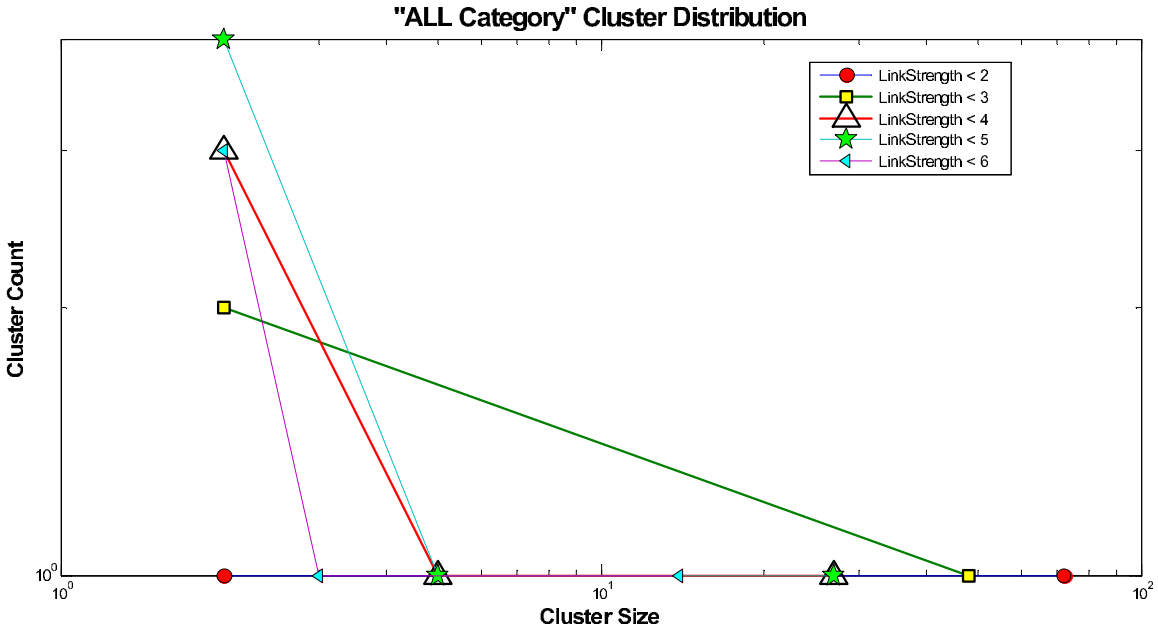

Figure 4. ALL label cluster frequency by cluster size per threshold value

To analyze this threshold behavior response, it is best with the aid of cluster distribution on figure 4 and following frequency graph in figure 5, which are graphical representations for the set of corresponding values contained in Table 1.

Table 2. Various statistics to compare clustering results for different threshold values for WisColl..

ALL-Categories	Number of clusters	Highest/ Lowest degree	Largest/ Smallest Cluster(size)	Coverage Total (pct.)	Coverage 1st / 2nd Cluster(pct.)
(>=3)	3	17 / 1	48 / 2	79.78	76.98 / 1.76
(>=4)	5	11 / 1	27 / 2	67.57	58.05 / 4.31
(>=5)	6	8 / 1	15 / 2	54.76	42.30 / 6.375
(>=7)	4	6 / 1	10 / 2	44.63	33.64 / 4.61
(>=10)	1	3 / 1	5 / —	21.67	21.67 / —

Figure 4 illustrates the distribution of clusters by cluster size and count for the *All-label* category for the clusters that remain after the links below a given threshold are removed. Each symbol marker denotes the threshold set of values, corresponding to those in Table 1. For each threshold, the process uncovers the number of clusters, through the combination of their sizes (as the number of linked nodes contained within the cluster), and the quantity of formed clusters for the specific size. To help illustrate, shown first is the case for threshold value of 3, for which all links below a value of 3 are removed, remaining only those with link-strength greater than 3 (i. e. $LS = 3$). This results in a network consisting of three clusters: two clusters, of size 2 each, as containing two nodes (right side marker), and one cluster of size 48. This sample case is included in figure 7 for the links of value of 3 or greater.

In Figure 5, a different perspective is provided for the same set of threshold values. For each threshold in Table 1, the number of clusters by size (i.e. number of nodes within the cluster) is presented, where the last column, when present, represents the number of clusters whose size is greater than 5. To this last aggregated column, a label is added for the size of the cluster that it represents.

From the graph, a couple of patterns can be identified. At first sight, a pattern for cluster size 2 quickly stands out. The pattern seems to suggest that as the process

increases the threshold, the number of the two node-pair size clusters increases, until an inflection point, after which their frequency decreases. The transition point, centered around 5, seems to coincide with the largest number of formed clusters. The second pattern is embedded in the aggregated column. In this pattern, the size for the largest cluster continuously decreases as the link strength value increases. The details for this trend can best be observed in Figure 6.

The significance of the data displayed for these initial results, more than the patterns or trends it unveils, is their support to the underlying concepts. As it has already been stated, most blogs are tagged with a pair of labels. Hence, the pattern observed by the size 2 clusters correlates with the notion that the two-node-pairs constitute the smallest cluster unit, and most susceptible to variances in link strength. There is some evidence of this dependency in the "second largest" component displayed in Figure 6, where large cluster nodes do not break into evenly sized clusters with subsequent thresholds, but rather spawns smaller size ones, suggesting blogs with closely related label pairs but displaying casual link to the main cluster, as the ones early breaking off.

This reasoning is further supported by the cluster *coverage* analysis results included in Table 2.

From the table, it can be observed that the clustering results for threshold=7 and 10 are quite poor. They both have very small number of clusters, i.e. 4 and 1 respectively, with very few nodes (as inferred by the corresponding highest degree statistics). Moreover, the coverage of the entire dataset for both these threshold values is very low.

When comparing clustering results for threshold values of 3, 4, and 5, it can be acknowledged that although having the maximum total coverage for threshold value

of 3, the resulting clustering is highly unbalanced, as evidenced by the coverage of the 1^{st} (biggest) cluster, which is almost equal to the total coverage, while the coverage of the second biggest, is just a fraction. For threshold value of 4 the total coverage is higher than that for 5, but the difference between 1^{st} and 2^{nd} cluster is higher for threshold=4 than the one for threshold=5.

This implies that the cluster size is much more unbalanced for threshold=4 than the one that is achieved for threshold=5. This is also confirmed by the largest cluster size values (given by the highest degree statistics) for both threshold=4 and threshold=5. Moreover, the search space reduction is less for the 1^{st} cluster for threshold=4 and since it is the 1^{st} cluster, many search queries will get their results from it. Hence a threshold=5 is set for the rest of the experiments for this data set composition. This also depicts the flexibility of the system. Depending on the domain, one can vary the threshold and obtain clusters of different size and different coverage.

Then is to explore a different venue to further validate an increase confidence in the threshold=5 selection. For this, refer to the clusters displayed in Figure 9 and Figure 10 that should help support the assessment derived from Table 2. The images illustrates how sparse these clusters are, the result of high threshold values for their respective link strengths, i.e. 7 and 10. Figure 8 exhibits the best clustering results among all other threshold values, further validating the selection of 5 as the optimal link strength threshold value.

The series of figures just referenced, those comprising Figure 7 through Figure 10 [6], provide a visual perspective into the analysis, where it can now include cluster visualization results for the All-label link category, for representative threshold values

[6]Pajek was used to create the visualizations.

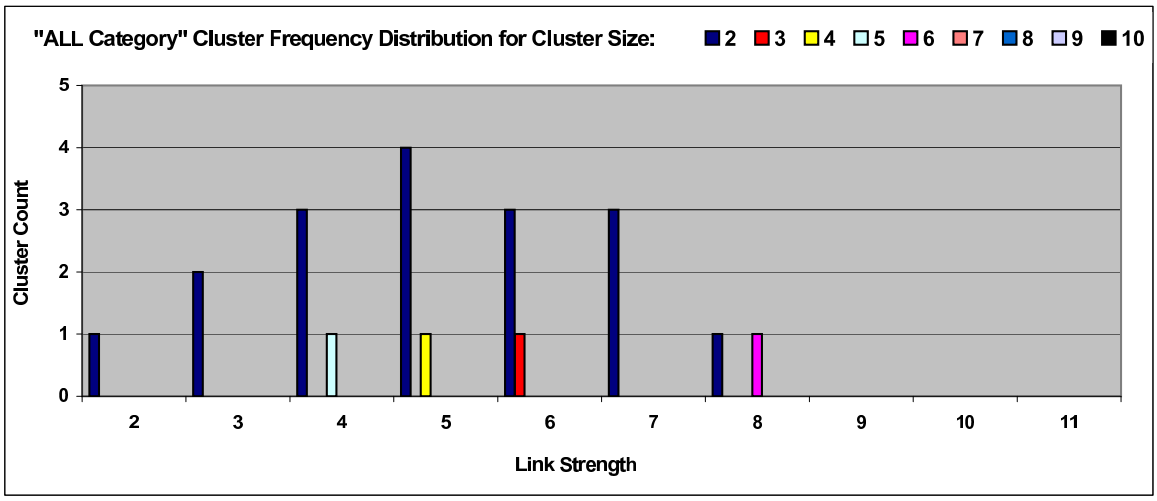

Figure 5. ALL label cluster histogram for small size clusters per threshold value

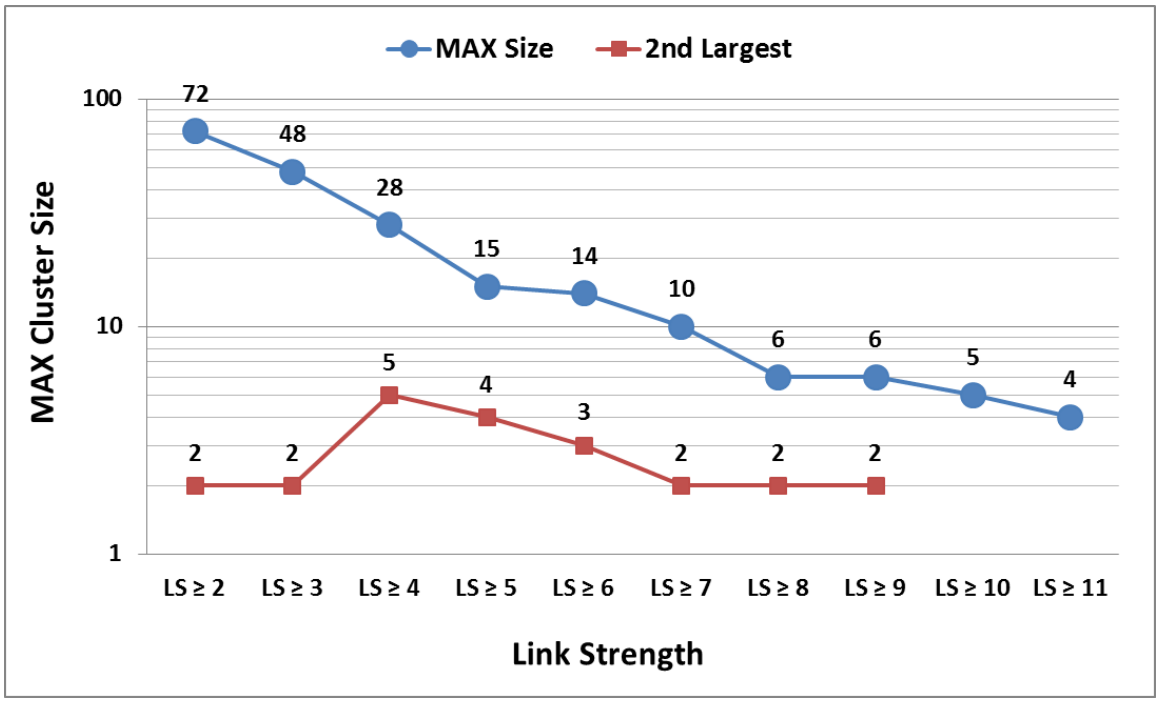

Figure 6. MAX cluster size per corresponding threshold value.

of 3, 5, 7 and 10. Contour lines are used to highlight and delimit the clusters. Node placement was maintained throughout the subsequent figures to facilitate visualize cluster's progression and transformation.

Notice that in these figures, Link Strength is denoted by the values on the edges. Names of the nodes depict the labels assigned by the bloggers to the blog sites. A cluster of labels would represent a cluster of all the blog sites that are labeled with one of these labels. Notice that various of the node's labels incorporate the use of >, like in the case of `Gaming>Computer & Video Games` (i.e. node in top-right cluster of Figure 7), to depict the hierarchical structure of the label. Here the blog sites are labeled `Computer & Video Games` which is a child of `Gaming`.

For threshold >= 3, total coverage is highest but notice it has a single large cluster and two very small clusters depicted by the cluster coverage. Similar is the case for threshold >= 4, 7, and 10. This indicates that highly unbalanced clusters are achieved at other thresholds as compared with threshold >= 5. This value coincides with the previous notion that this is the transition point as shown in Figure 4 and Figure 5. Hence a threshold=5 is set for the remainder of the validation experiments.

From the figures, it can be observed that for Figure 7, for threshold >= 3, though the total coverage is highest thanks to the large cluster, the generated clusters are highly unbalanced, exhibiting high size discrepancy among the collection of generated clusters for this threshold. This is as statistically recorded on Table 2. On a closer inspection, a visual scan of the labels composing the largest cluster suggests various sets of labels that could better be grouped as their own cluster, such as those related to Food and Drink (approximately top mid of largest cluster), and Technology (rightmost of largest cluster).

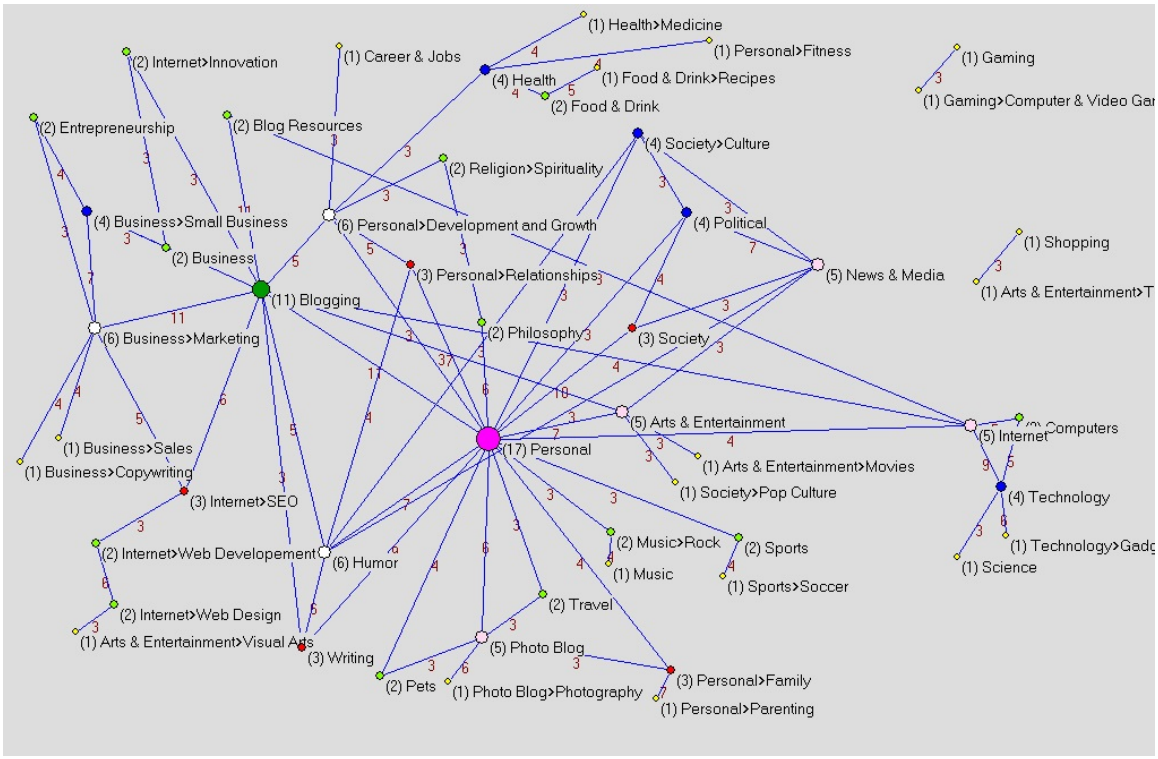

Figure 7. WisColl results for link strength ≥ 3 for All-label dataset.

These are precisely among the clusters that are unveiled when increasing the threshold to 4, and eventually 5, as illustrated Figure 8. It is worth noticing from the figures that for threshold >= 5, there is a better distribution of the clusters, and except for the largest, they are rather comparable, visually exhibiting greater balance relative to the other threshold value figures, once again, providing support to the decision of selecting threshold=5 for the validation experiments.

Similar projections can be derived for the clusters in Figure 8 to the ones that persist into Figure 9, though in this case, fewer clusters remain, resulting in less coverage.

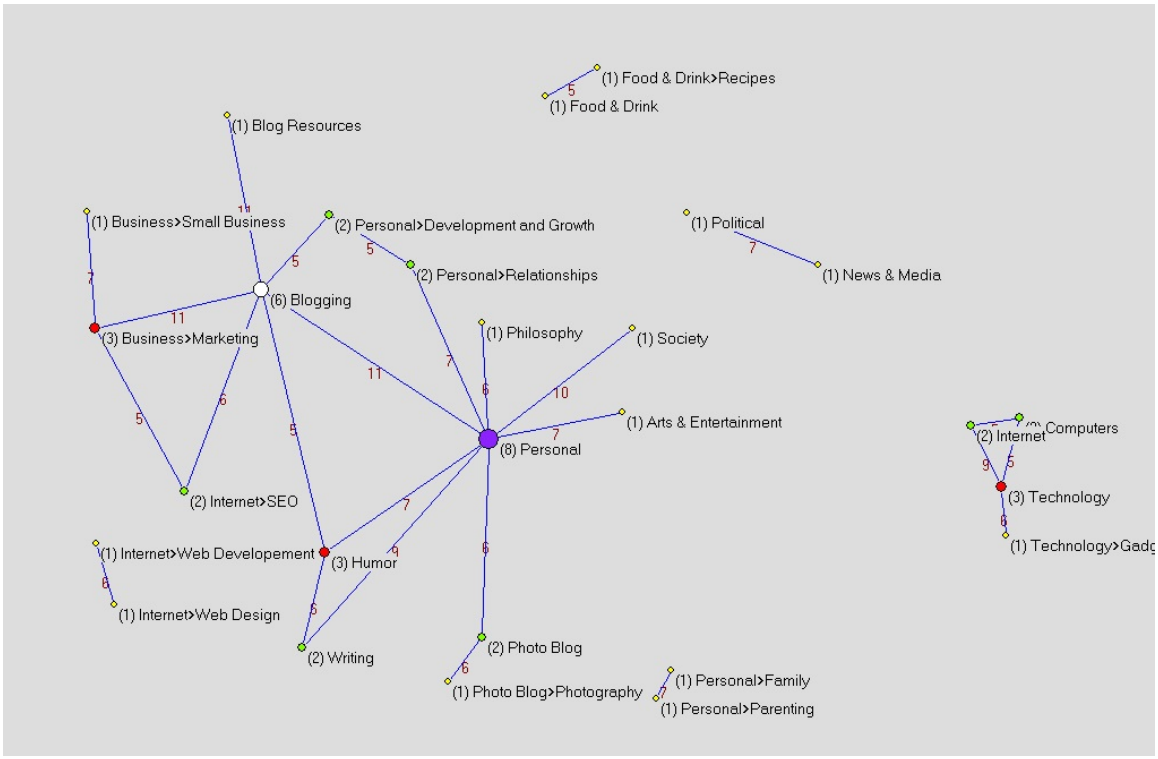

Figure 8. WisColl results for link strength ≥ 5 for All-label dataset.

4.3 Label Hierarchy

At this stage, the object is to study the effect of structural information of labels on WisColl. For this experiment is to consider all the three variants of the dataset, i.e., Top-level, All-label, and One node-split.

A sample of the clustering results for Top-level is shown in Figure 11. This option performs the worst among all the three variants. The cluster size is highly unbalanced. There is a single cluster to which all the nodes belong. This is the case for all threshold levels. No contour lines are drawn since no additional clusters are generated at any level. A sample clustering results for One node-split is shown in Figure 12. The results are a slightly better than Top-level but still the clustering quality is poor with unbalanced cluster size, as only in a couple of instances, more than a single cluster

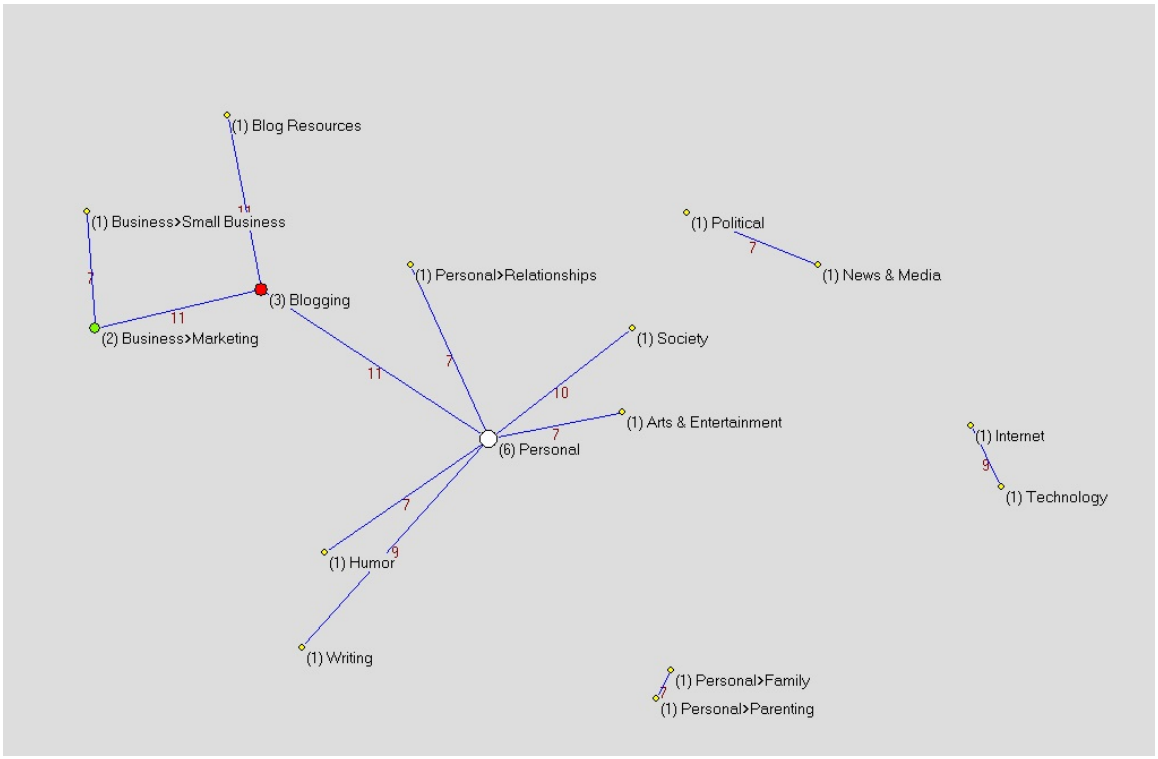

Figure 9. isColl results for link strength ≥ 7 for All-label dataset.

is present. As such, best clustering results are obtained with All-label as shown in Figures 7 through 9.

To compare, the statistics of clusters obtained from WisColl for different versions of datasets are contained in Table 3. Although the total coverage is maximum for Top-level label structure, there is only one cluster that connects all the labels. This results in 100% coverage for the 1^{st} cluster. So there is no search space reduction at all. Every time a query comes the results are returned from the 1^{st} and only cluster and since it contains all the labels, whole dataset needs to be searched. Similarly, results for One node-split show that the cluster size is highly unbalanced.

There are only 3 clusters with the 1^{st} cluster having majority of coverage (=76.44%) and the difference between 1^{st} and 2^{nd} cluster is very small. This largely affects the search space reduction. Results for All-labels has the lowest coverage but the cluster

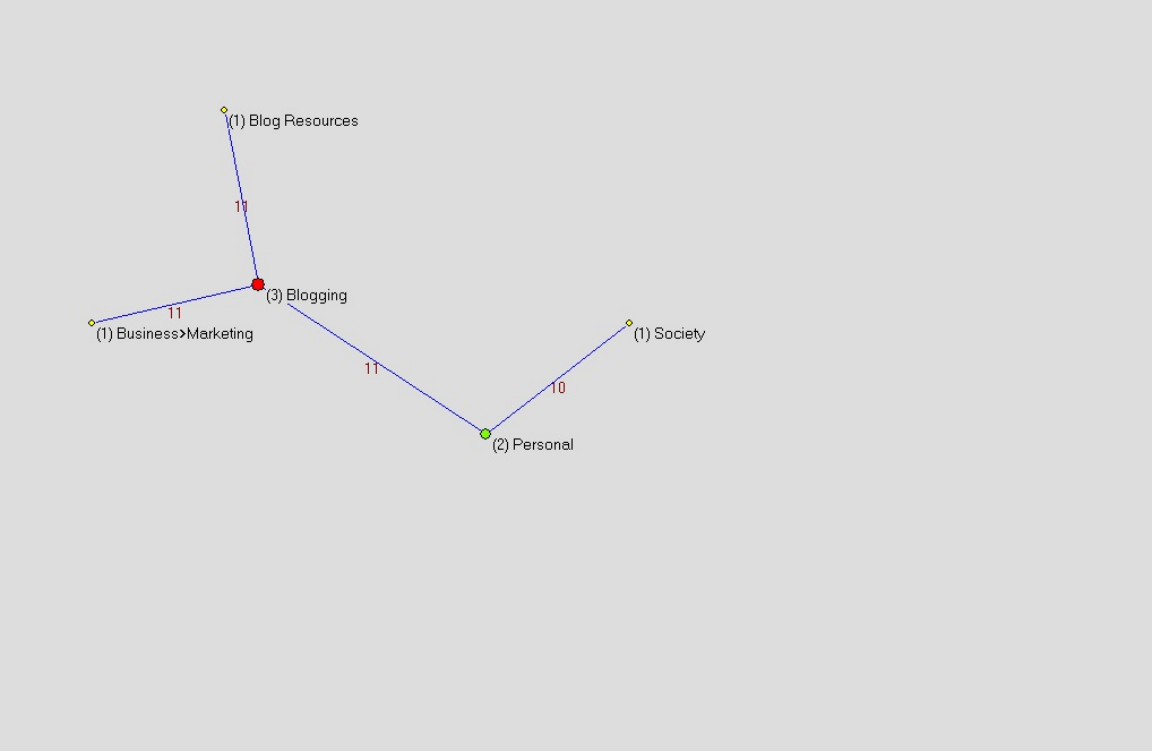

Figure 10. WisColl results for link strength ≥ 10 for All-label.

Category	Number of clusters	Highest degree	Lowest degree	Largest cluster size	Smallest cluster size	Coverage Total %	Coverage 1st cluster %	Coverage 2nd cluster %
All-categories, >= 5	6	8	1	15	2	54.76	42.3	6.375
Top-level	1	16	1	22		100	100	
One node-split	3	9	1	21	2	82.87	76.44	3.88

Table 3. Various statistics to compare clustering results for different label structure for WisColl.

sizes are not as unbalanced. Moreover, the difference between the coverage for 1^{st} and 2^{nd} clusters is larger than One node-split. This leads to better search space reduction. This shows that leveraging the complete structure of collective wisdom gives best results as compared to exploiting a part of it. This validates that the more collective wisdom is available the better it is.

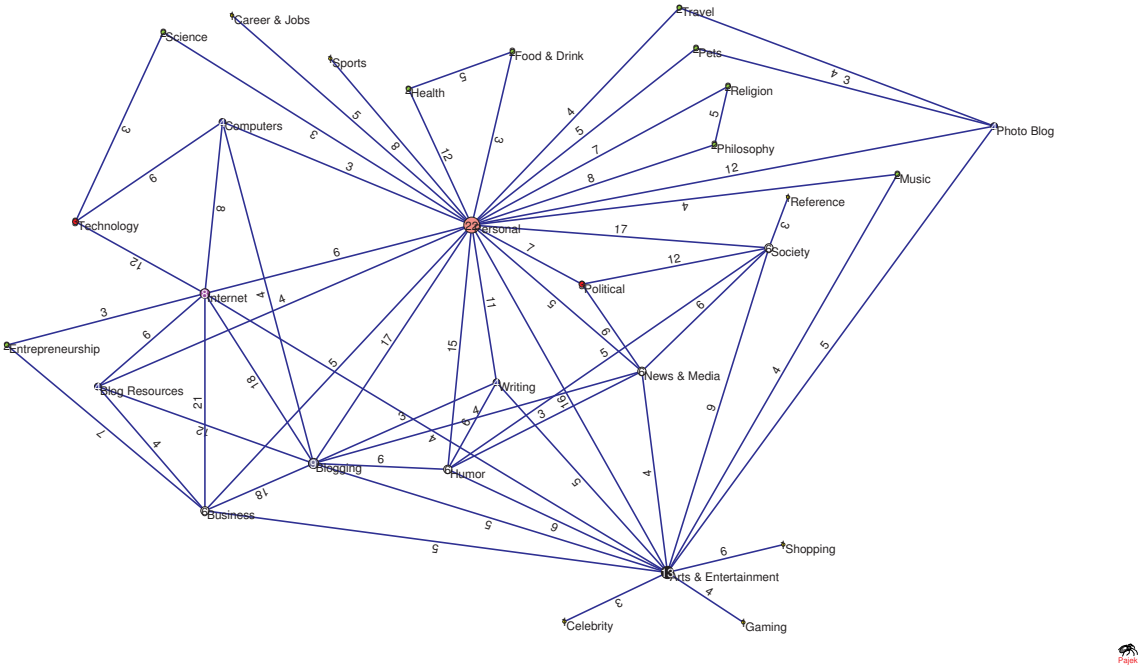

Figure 11. WisColl results for link strength ≥ 3 for Top-Level Label dataset.

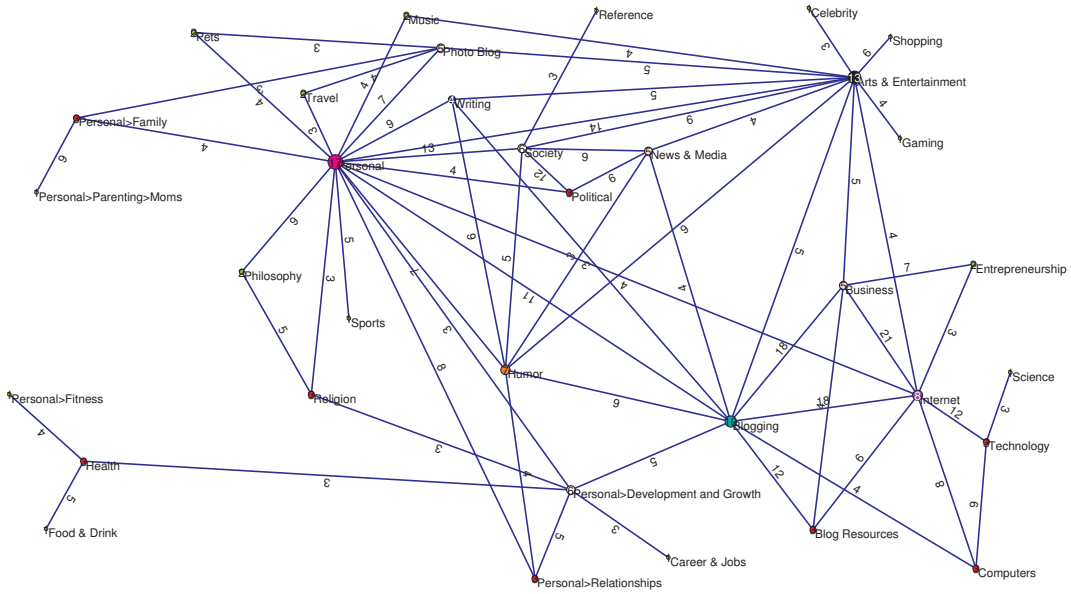

Figure 12. WisColl results for link strength ≥ 3 for Personal Label dataset.

4.4 Visualizations - Pajek

Here is to visualize the results of WisColl algorithm (e.g. Figure 8) and baseline algorithm to study the advantages of collective wisdom. Results for the baseline clustering were generated in an analogous fashion as to what was collected for the WisColl approach. As such, first is to study the Baseline's link strengths range of values. These values are shown in Table 4. The table shows the range of values, and their distribution, for all of the $346,921$ edge line pair values, corresponding to 842 Baseline nodes.

Figure 13 illustrates the dependency of threshold value, and the distribution of edge-connected clusters by size and frequency, for the Baseline range of threshold values, corresponding to those values in Table 4. This figure shows a similar behavior as was observed for Figure 4, that as the process increases the threshold, the number of smaller size clusters increases, and the larger size clusters decreases, up to a certain transition point, after which the total number of remaining nodes in the network decreases. For the Baseline case, this transition point is centered around a threshold value of 0.75 to 0.80. This is illustrated by Figure 14 which shows cluster size distribution and the transition, with increasing threshold value, for the smaller size clusters. Figure 15 presents the edge-connected cluster visualization results for the Baseline link strength range, for representative threshold value of 0.80

Best visualization of cluster type clustering forming results for baseline approach were achieved with the threshold $\tau = 0.9$ for Figure 16 . Here nodes represent the blog sites or bloggers. For easier comparison it is best to also display the labels of their blog sites besides their name. For example, a node label like, `emom=Small Business:Moms`, indicates that the blogger `emom` has a blog site with labels `Small Business` and `Moms`.

Cluster quality for both the approaches could be compared by looking at the labels of the cluster members. However, the label information is not used while clustering in baseline approach. What follows is the report for the differences between the two approaches based on the results as follows:

1. There are too many clusters obtained from baseline approach and many have very small size (most of them are 2-member clusters). However, this is not the case with WisColl.

2. As a result of too many small sized clusters, clusters are too focussed. This affects the insertions of new blog site later on. Cluster configurations are highly unstable in such a focussed clustering. For example `cozimono` = `Music:Rock:Pop` and `billiam` = `Music:Rock:Pop` are clustered together. This group is highly focussed and if a new blog about Music comes is added then it won't be assigned to this group.

3. Deeper analysis shows that some clusters obtained from baseline clustering, have members whose blog site labels are not semantically related. For example, `bluemonkey jammies` = `Humor:Personal` and `emperoranton` = `SEO: Marketing` are clustered together. However, the labels are totally different and are not at all semantically related. There are several such clusters obtained from baseline clustering approach. This shows that baseline clustering does not give semantically coherent clusters. This is because vector space clustering using blog posts are susceptible to text noise, and blogs are usually noisy. Also blogs are dynamic in nature with the blogger occasionally posting about different topics. Such off topic posts affect the clustering using vector space methods. However WisColl gives high-quality, semantically coherent clusters. For example, clusters having members like `Internet> Web Design` and `Internet> Web Development`; `Food & Drink` and `Food & Drink> Recipes`; `Internet`, `Computers`, `Technology`, and `Technology> Gadgets` etc. are semantically related.

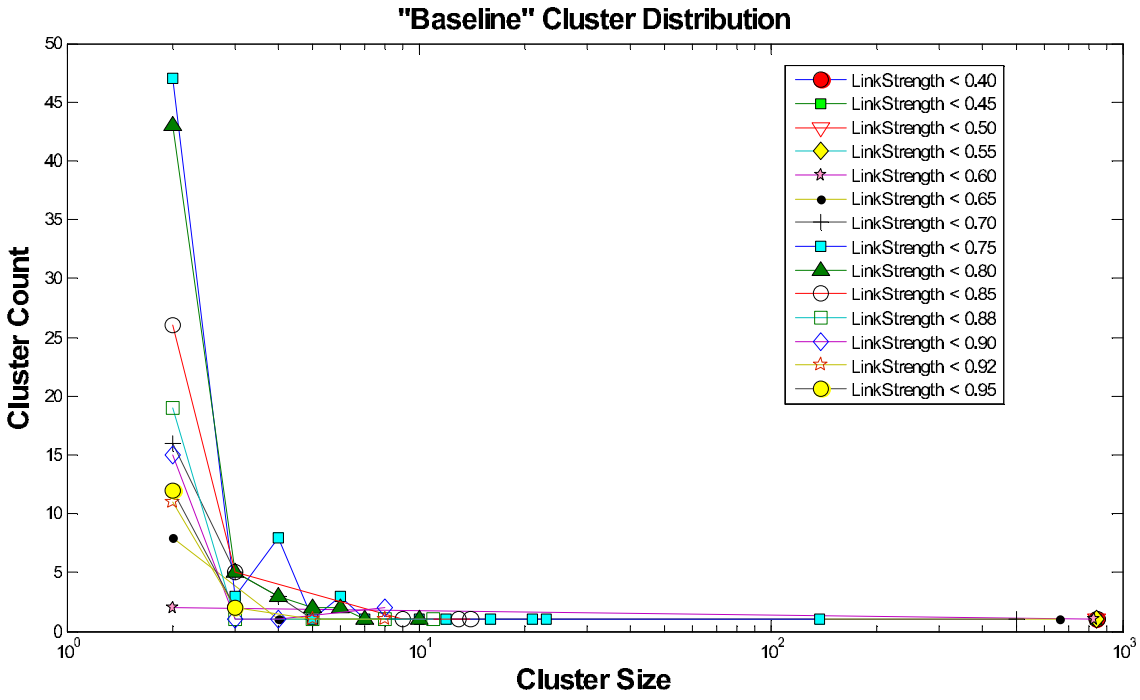

Figure 13. Baseline cluster frequency by cluster size per threshold value

4. Several clusters obtained from baseline approach have members that have exactly same labels. For example, the cluster with bloggers `emom` and `geraelindsey` have the same labels, i.e., `Small Business` and `Moms`. This does not help in identifying relationships between blog sites that have different themes. Clustering blog sites that have different yet related theme/topics are more helpful. WisColl generates clusters that have blog sites with topics like, `Technology`, `Computers`, `Internet`, and `Technology> Gadgets`. Such a cluster serves a better purpose for various applications like search, organization of information, etc.

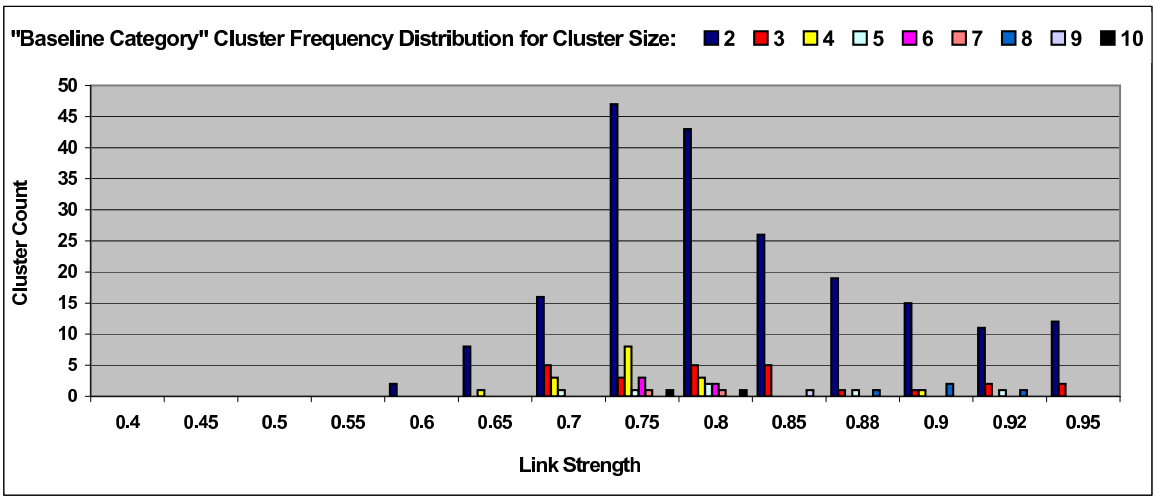

Figure 14. Baseline cluster histogram for small size clusters per threshold value

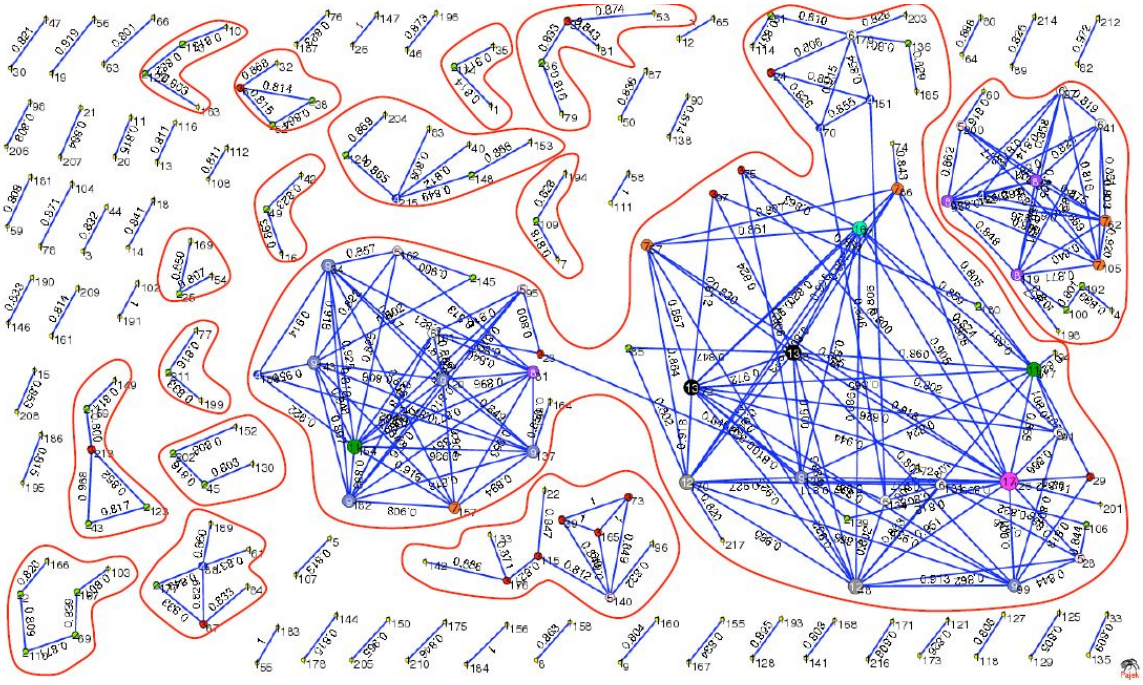

Figure 15. Results for link strength ≥ 0.80 for Baseline dataset.

Table 4. Baseline Link Strength statistics.

Line Value Range	Freq	Freq(pct.)	CumFreq	CumFreq(pct.)
0.0000 ... 0.0417	53288	15.3603	53288	15.3603
0.0417 ... 0.0833	51239	14.7696	104527	30.1299
0.0833 ... 0.1250	47085	13.5723	151612	43.7022
0.1250 ... 0.1667	40021	11.5361	191633	55.2382
0.1667 ... 0.2083	33588	9.6817	225221	64.9200
0.2083 ... 0.2500	27705	7.9860	252926	72.9059
0.2500 ... 0.2917	22250	6.4136	275176	79.3195
0.2917 ... 0.3333	18505	5.3341	293681	84.6536
0.3333 ... 0.3750	14759	4.2543	308440	88.9078
0.3750 ... 0.4167	11054	3.1863	319494	92.0942
0.4167 ... 0.4583	8581	2.4735	328075	94.5676
0.4583 ... 0.5000	6052	1.7445	334127	96.3121
0.5000 ... 0.5417	4290	1.2366	338417	97.5487
0.5417 ... 0.5833	2823	0.8137	341240	98.3625
0.5833 ... 0.6250	2382	0.6866	343622	99.0491
0.6250 ... 0.6667	1364	0.3932	344986	99.4422
0.6667 ... 0.7083	851	0.2453	345837	99.6875
0.7083 ... 0.7500	488	0.1407	346325	99.8282
0.7500 ... 0.7917	279	0.0804	346604	99.9086
0.7917 ... 0.8333	167	0.0481	346771	99.9568
0.8333 ... 0.8750	79	0.0228	346850	99.9795
0.8750 ... 0.9167	33	0.0095	346883	99.9890
0.9167 ... 0.9583	23	0.0066	346906	99.9957
0.9583 ... 1.0000	15	0.0043	346921	100.0000
Totals	346921	100.0000		

4.5 K-means vs. Hierarchical Results

A Hierarchical clustering was generated for the 27 labels identified by WisColl for Link Strength 5 or greater, (see Table 5). The clustering was achieved using Pajek's Hierarchical Ward method, and is illustrated in Figure 17. From the hierarchical diagram, 7 major clusters were selected, as illustrated by Table 5. From this, it was then generated a K-Means clustering for k=7 for the labels to analyze how well K-Means and Hierarchical clustering compared between the two methods.

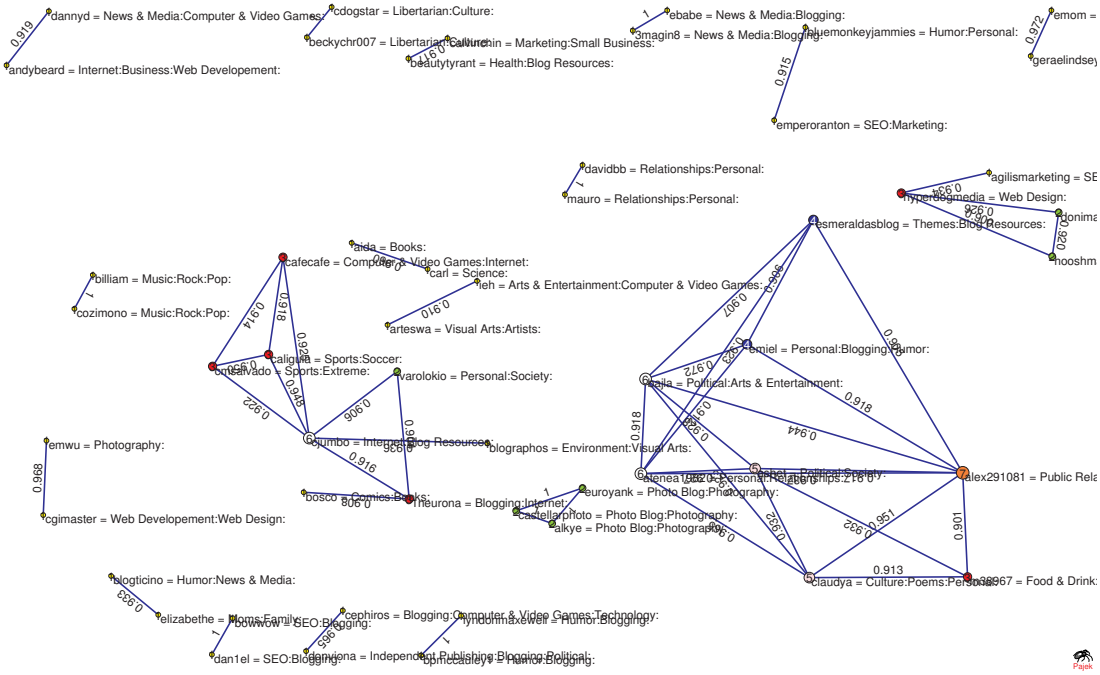

Figure 16. Results for link strength ≥ 0.90 for Baseline, labels included

In order to compare how WisColl based clustering fares with regards to the baseline blogger clustering, similar Hierarchical and K-Means clusters for the baseline's blogger space was then generated as follows. In order to be able to compare between the results obtained from the label space, into the blogger space, each vector in the label space was "mapped" to its corresponding vector(s) in the blogger space. This was accomplished by associating the bloggers to each of the labels in the label cluster, given the blogger had used that label in his/her blog. This transforms the clusters in the label space to an equivalent cluster made out of bloggers in the blogger space. The mapping was generated for both Hierarchical and K-Means results. Since many of the bloggers had used more than one label for a blog, if the between-cluster distance is computed using single-link, then many of these distances would be 0. This will skew the distribution of the cluster distances. Therefore the task is to then calculate the

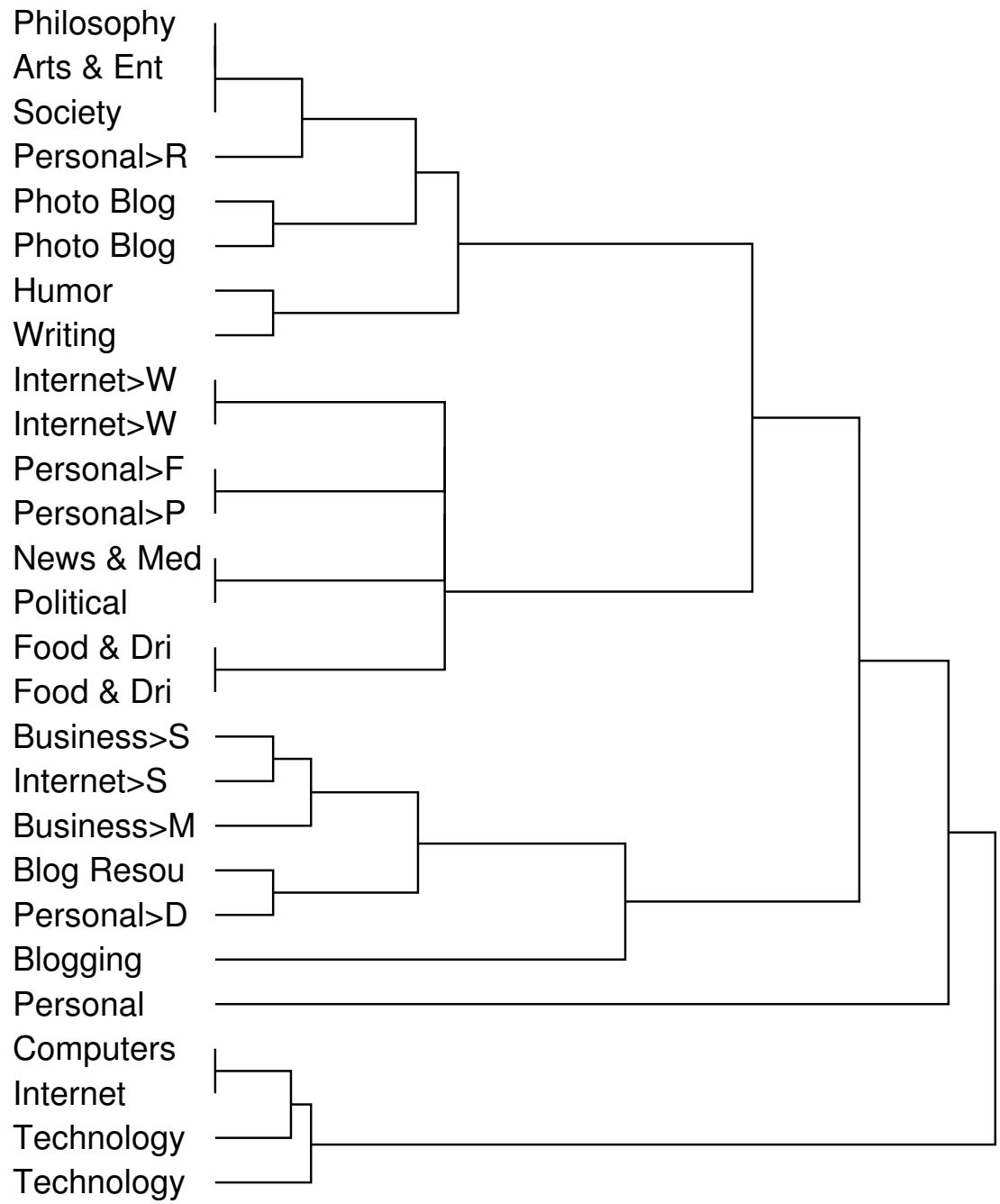

Figure 17. Hierarchical clustering for Link Strength ≥ 5 for All-label set

Table 5. Hierarchical clustering cluster assignment for Link Strength ≥ 5 for All-label

Index	Clus ID	Cluster Component
1	1	Philosophy
2	1	Arts-n-Entertainment
3	1	Society
4	1	Personal>Relationships
5	1	Photo Blog
6	1	Photo Blog>Photography
7	2	Humor
8	2	Writing
9	3	Internet>Web Design
10	3	Internet>Web Development
11	3	Personal>Family
12	3	Personal>Parenting
13	3	News-n-Media
14	3	Political
15	3	Food-n-Drink
16	3	Food-n-Drink>Recipes
17	4	Business>Small Business
18	4	Internet>SEO
19	4	Business>Marketing
20	4	Blog Resources
21	4	Personal>Development and Growth
22	5	Blogging
23	6	Personal
24	7	Computers
25	7	Internet
26	7	Technology>Gadgets
27	7	Technology

between-cluster and within-cluster distances using the average-link.

The mapping generates same number of clusters in the blogger space as the number of clusters obtained in label space. Since selected 7 clusters for hierarchical clustering in the label space, after mapping obtained 7 clusters of bloggers.

Next is to try to observe the best value for k in k-means. For the analysis and

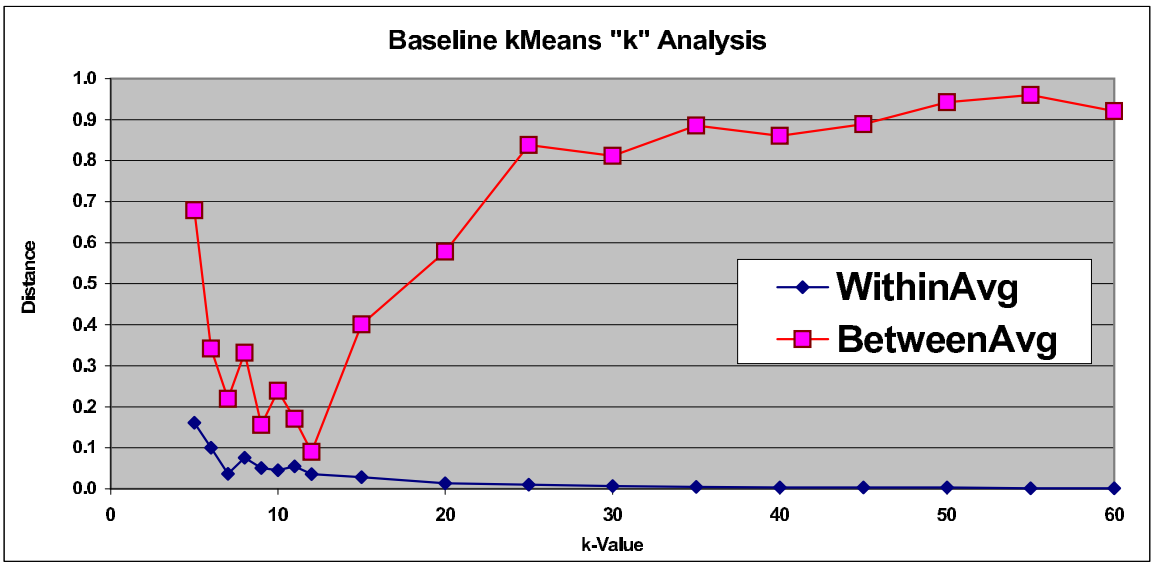

Figure 18. kMeans k-Analysis for Baseline dataset.

comparison, following the premise discussed in the link-based clustering section, where it was presented the min and max criteria based on Formulation 3.1, which minimizes the within-cluster distance between the cluster members (i.e., more cohesive clusters) and Formulation 3.2 with regards to maximizing the between-cluster distance (i.e., well separated clusters). Figure 18 shows these results for the blogger clusters. To asses the value of k, is to take advantage from the results obtained from Figure 13, which suggests that based on link-strength analysis, the highest cluster count is for link strength 0.75, which is slightly above 60. Hence, is to use 60 as the maximum value of k. Then is to perform clustering of bloggers for k=5 to 60. As shown in the figure, can observe a lot of fluctuations in the range k=5 to k=12. Upon deeper analysis it shows that can obtain lowest within-cluster distance for k=7. This results is in accordance to the hierarchical clustering as it also generated 7 clusters. Note that for k>25 even though it looks that the clustering is better (large between cluster distance and small within cluster distance) but in reality results in obtaining a significant number of 2 member clusters for higher value of k, just like in Figure 15.

The results comparing Hierarchical and K-Means clusters are summarized in Table 6. The table shows three categories: a) "Baseline - Blogger Space" refers to the clustering generated by clustering in the blogger's space; b) "WisColl - label Space", evaluates the clustering generated using label relation graph in the label's space; and c) In "WisColl - Blogger Space" transformed the label space clusters into the blogger space as described previously. It is in this last case, since it projects the labels into the bloggers space, it then allows make a fair comparison between the methods.

From the results in Table 6, two observations can be made. First, WisColl performs better than Baseline in both clustering algorithms, k-means and hierarchical (i.e., categories 'a' and 'c' in the table). Although can get lower within-cluster distance for baseline approach than the WisColl, the between-cluster distance is higher for baseline as compared to WisColl. This means that though can have cohesive clusters from baseline approach, they are not well separated. Also, the variance in the between-cluster distance of baseline approach is much higher than that of WisColl. Note that this comparison is made in blogger space to be fair. Second, k-means performs better than hierarchical for WisColl in label space, since k-means has lower within-cluster and higher between-cluster distance. However, in blogger space for WisColl, hierarchical clustering performs a little better than k-means clustering with higher between-cluster distance and comparable within-cluster distance.

To summarize the results, it can clearly be seen that this method, when compared with baseline after mapping the labels to blogger space, has a better "separation" and tighter variation between the clusters.

Table 6. WisColl vs. Baseline approach using k-Means and Hierarchical clustering.

Type	Method	Within	Between
2*(a) Baseline-Blogger space	k-Means	0.0363 ± 0.1264	0.2194 ± 0.1301
	Hierarchical	0.0890 ± 0.1186	0.3644 ± 0.0903
2*(b) WisColl-label space	k-Means	0.0615 ± 0.1643	0.2860 ± 0.0536
	Hierarchical	0.0857 ± 0.1672	0.2761 ± 0.0571
2*(c) WisColl-Blogger space	k-Means	0.0844 ± 0.0995	0.7090 ± 0.0143
	Hierarchical	0.0849 ± 0.0943	0.8118 ± 0.0047

Chapter 5

MULTI-LABEL AGGREGATION

5.1 Introduction: Troubles in Label-land

Thus far have demonstrated the benefits and effectiveness in generating meaningful clusters that can be drawn from the collective wisdom garnered by exploiting the blogs' labels and content, and to whom have come to refer as WisColl. However, one consideration jumps into mind whenever clustering, classifying, or alike. These results can be very dependent on the accuracy, consistency and reliability of the defining parameters. Appropriately for this case, the labeled tags that are entrusted to represent a blogsite. This is even more aggravated when a multi-labeled environment is at stake, as the opportunity of mislabeling is multiplied by a factor equal to the number of labels associated for a particular post, for case in point.

Labeling of data is a very expensive and tedious affair. In most classification environments, particularly for text mining, labeled categorization is by far disproportionally outweighed by the amount of unlabeled data. But when it does happen, that is, labeled textual data, class mislabeling can quickly become a nuisance, which can contribute to levels of noise into the results. Considering that labeling is a human treat, must always entertain into the calculations the possibility that some level of mislabeling might be expected. The source of mislabeling can be various, such as a simple entry error, to a more complex or profound nature, such as interpretation and subjectivity, or even locality or cultural upbringing, where either the subject to be labeled could support a variety of fitting descriptors, or quite the contrary, has a high

level of ambiguity that no amount of labels would seem to qualify. A comprehensive list of such sources are detailed in (Frénay and Verleysen 2014).

Trying to overcome this concern has been the source of wide area of research. As a result, solutions and alternatives abound throughout the research community, devising methods and techniques that will attempt identify and eliminate these instances. Such work can be outlined with early contributions by (Brodley and Friedl 1999), which first details the sources as described earlier, and try to devise a filtering algorithm by trying to identify outliers within a set. This is narrowed by (Sun et al. 2009), which simplifies using a Bayes classifier at the instance level, removing those with the lowest information entropy with respect to the pool of labels being considered. Almost analogous is (Esuli and 0001 2013) label ranking technique devised to identify the likeliness that an instance might be misclassified. However distinct and varied are the sampled approaches presented, one common theme among these sampled works, and throughout the mining community, is that they all agree that mislabeling can lead to inefficient and inaccurate classification results, that should be addressed, and taken into consideration when evaluating results.

It is here that the case is made that the enforcement of the WisColl process can overcome some of these labeling maladies. In following the strategy, the imposing of the strongest links philosophy when selecting the labels, skirts off levels of mislabeling associated with spurious labeling noise, thus reducing the chances of being swayed that could impact the process and results. In supporting this, seems to agree with (Frénay and Verleysen 2014) in which it is assessed that 'the incorrectness of a label is assumed, when no other information is available'. Which can here be interpreted as labels lacking supporting evidence, or associated context, might need to be revisited.

This leads to boldly infer that the opposite should also be equally true. That is,

that the correctness of a label should be assumed when an abundance of information is available, as in this case, through the support of the collective wisdom. This condition quite fits the case when considering the strength brought on by the collective wisdom increasing the level of confidence in the ability attributed to the bloggers when making their labeling assessment in their effort to best portray their product creation that contains their blogs.

However, not to belittle the concern of noise in the label data just described, best consider a more challenging proposal in the form of labels that may provide little meaning, or support to the contents of a blog. To no one's surprise, this work uncovers that a number of bloggers are not as zealous when labeling their blogs, thus introducing a different type of noise variability. This is attested by the lack of refined categories used for the labeling applied.

Though somewhat suspected, analysis suggests bloggers for these cases may have opted instead for run-of-the-mill tags, or may just be simply an act of indifference, in which when faced with too many potential label options, more likely will be the case that the blogger does not find a categorizing label that perfectly fits. As such, they may simply may rush through the labeling process, opting perhaps with a popular, or a broad catch all category that will just suffice without much ceremony. Another equally disrupting behavior is the one in which bloggers launch posts whose contents stray from the descriptors originally implemented to label the blog they are posting from, often referred as the *topic irregularity problem* woes. This type of behavior can further be extended to consider those blogs that, as they evolve in time, depart from their original intent and purpose, failing in the process to update or refresh their categorization into one that more accurately reflects their shift in focus.

Unfortunately, these cases illustrates that the role of the labels or tags may not be well understood among some in the blogging community. Care should be taken to avoid categorizing with labels that may be too generic or may loosely represent their contents. Failing to exercise due diligence in a fitting tag selection, misses the opportunity to maximize, and take advantage, of the tag's full potential. Inaccurate tagging, whether unintentional or intentional, pays little justice to the blogger's postings, their potential followers, and advertisers.

Selecting the proper tag to associate to a posting may prove critical into categorization along other similar postings. This is particularly significant when considering that one of the blogger's main objectives when posting their material should be to connect with others and be influential, and such demeanor would certainly be detrimental towards achieving this goal, risking as such, the potential for discovery, recognition, and their influence upon others. It also puts a much hamper on the work of search engines and categorization.

Hence, understanding the source of mislabeling can help compensate and reduce dependency on perfect labeling. Hence, it is best to seek understand the nature and consequences of the mislabeling basis. Being that labeling and tagging are human trades, in the exploration for this work, also include some of the philosophy and psychology characteristic elements of tags. Even for instances in which the labels, or tags, are generated by automated means, they are still meant for human consumption, should try understand.

There is quite a lot at stake when considering a label or tag. The psychological effect of choosing a best fitting tag can be both daunting and a very challenging affair, particularly when the blogger has to take into account its potential effect with regards to search engines, and in the attracting of followers and advertisers. Selecting the best

label, should be carefully crafted, considering in the process the target audience, the message to convey, ensuring it synthesizes and captures the essence of the posting. In fact, tags and labels become *artefact of meaning* as per defined by (Halpin, Robu, and Shepherd 2007) for the reality of the blogger it is trying to project.

This work demonstrates the significance of proper labeling with a simple analogy anyone who has ever written a résumé or curriculum-vitae can quickly relate. In a résumé, or vitae, a few lines in a handful of pages may attempt summarize an individual's lifetime professional and academic achievements and accomplishments, in a manner that should convey interest, and signify a chance for a coveted position opportunity as a top candidate. As such, tremendous conscious care is incorporated into its crafting, and for many, a significant amount of anxiety, as the consequences, and rewards, could last a lifetime. Similarly for a post, the blogger is analogously reduced to a handful of carefully chosen key token words, equally seeking for that connection, or remuneration, as a favorite blog among other possible blogs with similar content, and though the level of anxiety might be at a lesser extent, the element is nevertheless, still present.

Blogs are yet to reach celebrity status, otherwise, a single label or tag should uniquely provide enough context to categorize them, which would greatly simplify and solve the blog's résumé analogy illustrated above. Consequently, as contents in the internet get more diverse and rich, a single label could not possibly do justice to its contents, particularly, in regards to search engines crawling throughout the web. There are indeed certain type of labels that could themselves suffice in determining a whole genre, or class category, but contrary to certain celebrities, or personalities, which can be defined by a mononym (i.e. 'Cleopatra', 'Curie', 'Einstein', 'JLo', 'Madonna',

'Monet', 'Picasso' or 'Spock') most blogs must rely on greater than one tag or label to properly discern among themselves.

Hence, bloggers must rely on a multi-labeled environment that would facilitate their finding by search engines when homing in their search and scope. But as such, they embark into a labeling conundrum or venture into a folksonomies space riddled with tags they must navigate. Though folksonomies do enrich the web, empowering the users with tags they see fit, creating their own vocabulary or jargon, which would be particularly most beneficial if observed within a close blogging community. However, the research community acknowledges its shortcomings (Hayes, Avesani, and Veeramachaneni 2007) (Halpin, Robu, and Shepherd 2007) (Mathes 2004), where its free-form permeates no pre-defined structure, except for one if imposed ad-hoc by the blogger. The blogger is do encouraged to tag the posts, and sometimes enticed with a tag map or cloud, but is under no obligation to comply, nor is form of enforcement applied. As such, problems abound such as ambiguity and synonyms control, with no apparent connection or link between the tags. Nevertheless, taxonomy schemes proposals abound, which favors the search engines when exploring the space, but serves little to no purpose to the blogger while at the task.

On the other hand, when structure is provided, such as this work through Blogcatalog, where a label hierarchy is most likely enforced, is not without its share of tribulations. Similar to taxonomies offered in the folksonomy space, a common framework is encountered in both approaches: a broad higher level category is first identified, followed with lesser level, but related, refined subcategories. However, this work seems to suggest that frequently the top broad category may be too broad, or simply gets on the way of the blogger who is seeking for a more defining label embedded deeper layer in the structure. For the blogger, it may feel like the proverbial

fitting a square peg into a round hole syndrome, in which the blogger needs to keep hammering until the right hole is found. On the other hand, if the initial 'broad' categories are too broad, they could be in fact, less informative, less discriminative, not very descriptive, and harder to discern.

This work uses the cases described above to help illustrate some of the contributing factors, and as such, the challenges imposed in regards to the accuracy, consistency and reliability of the tags that normally rely in the assertions to represent a blog. While there might be other contributors, the objective of this work is not into the discovery, or understanding of them all, but rather try to compensate and overcome them, limiting any potential adverse impact.

One such case, which will be used to help illustrate this work's strategy, becomes palpable when evaluating the graphic for the top labels addressed in the previous chapters as shown in Figure 19. For this figure, a tag such as `Personal` is quickly highlighted as from among the top labeling categories. This tag is no strange in this work, as it is one to which have made earlier remarks for its notable skewed distribution.

One of the several drawbacks of the `Personal` tag, is that of as a catch-all generic categorization. For starters, the label by itself can be interpreted in several ways, such as either as an adjective, referring to an individual, or as a noun, as for example, a section in a newspaper. As a standalone label, it offers no valuable information, lacking significant resolution to discern between blogs for further in-depth clustering. On pondering the popularity nature of the tag, should speculate it is perhaps misinterpreted at labeling stage as implying a possessive nature, that is, as "my personal blog". This work seek to alleviate this condition.

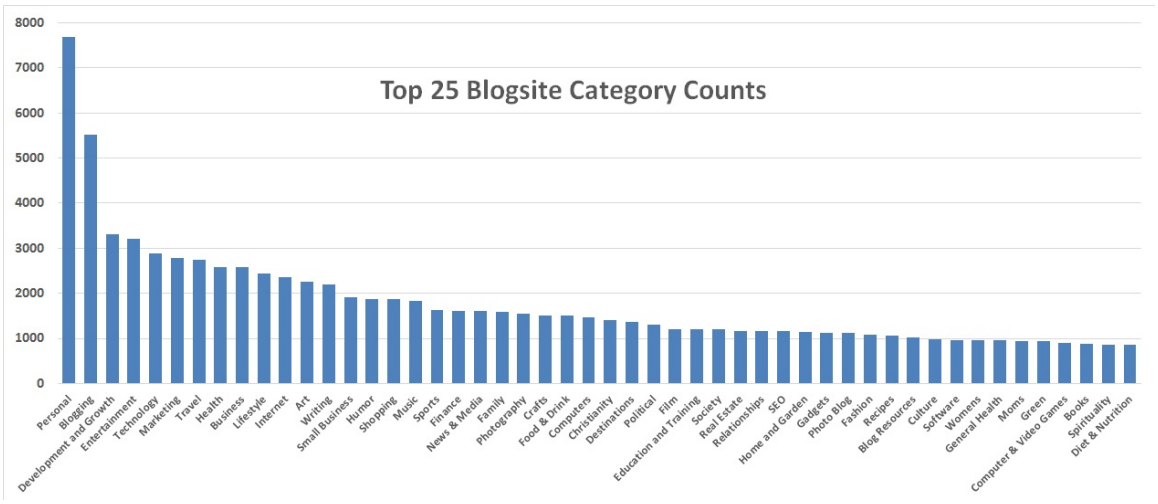

Figure 19. Top 25 Blogsites Distribution

This work aim to replace, or complement, `Personal` in a category pair with a more suitable and descriptive label, but with the imposed restriction that the replacement, or rather complement, not arbitrarily selected, but should be from among the set of categories who have exhibited a pre-existing relation to `Personal`. Thus delving on the premise that the blogger exercised greater care when selecting the pairing category to `Personal` for the blogsite, and trust that this second category closely represents the contents for the blog. Thus, the candidate category should also have a proven relation with the second category of the `Personal` category duo and should also closely typify the contents for the blogsite it will now co-represent.

To aid in this quest, at disposition is a collection consisting of little less than 110,000 blog records. Collected in the fashion as initially described, gathered the data by crawling the social network of four BlogCatalog bloggers who were selected from among the bloggers associated with labels containing the largest number of blog sites.

Since the intent in this process is to simply replace `Personal` with a plausible category for a category pair, to simplify and ensure that it initially limits the study to include blogsites with exactly two categories assigned to them. There are approximately

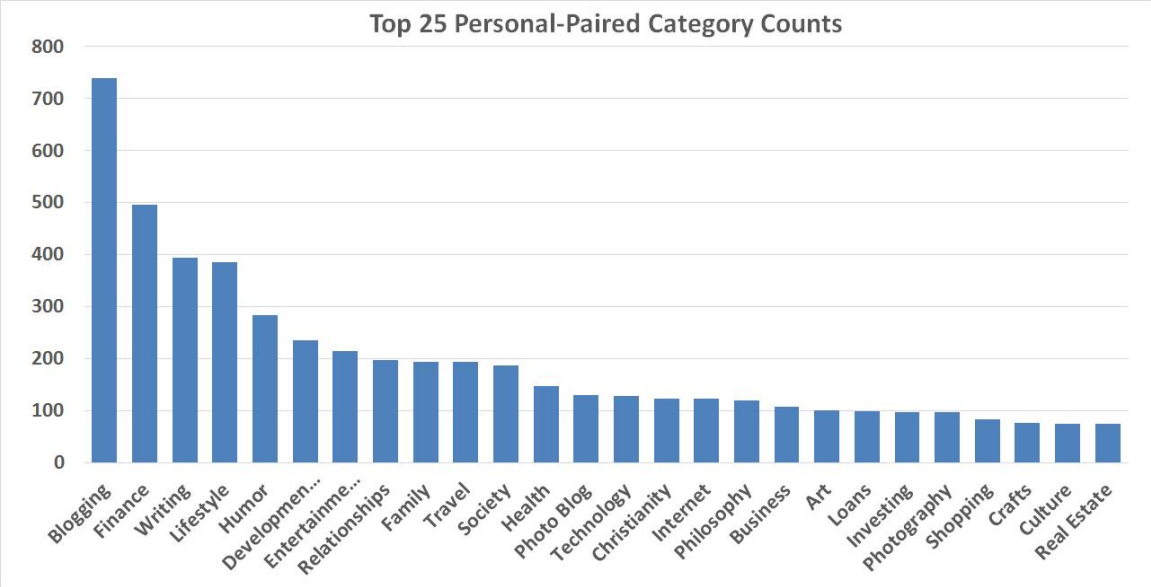

Figure 20. Top 25 Personal Category Pairs Distribution

1,000 records removed from the collection that exhibit greater than two categories. Also removed are those blogs with less than two category labels (approximately 15,000 records). From the remainder collection, also removed those blog sites which contain no post data (approximately 11,000 records), removed those blogs that had no associated tags (approximately 4,000 records) and removed those which contained small posts, that is, 50 words or less (approximately 6,000 records).

The aforementioned adjustments resulted in a net collection consisting of precisely 71,258 blog sites, from which was divided into two main camps: those that contain `Personal` as one of the pairing categories, such as, `Personal`-and-`Finance`, or `Personal`-and-`Writing` and the camp does not include `Personal` as one of the pairing categories, such as `Finance`-and-`Investing` or `Writing`-and-`Books`. Hence, there were identified 7,688 records that contained the `Personal` label as one of the category-pairs, with a remainder of 63,570 blog sites which did not.

From the Personal-inclusive group, were identified a total of 241 category labels, of which the distribution of the Top-25 Personal pairing categories can be observed in Figure 20. Thus can be gathered from the figure that Blogging, Finance, Writing and Lifestyle are among the top categories associated with Personal. The aim is then to explore the possibilities of finding more suitable and descriptive categories to replace Personal for the blog sites containing these top four categories. Since these seem to be the most popular label categories, chances are that the behavior should persist when exploring on the non-inclusive Personal group. As such, the attempt is to find reasonable fitting replacements to the Personal category among the pairing categories associated with Blogging, Finance, Writing and Lifestyle from content-equivalent blogs within the non-Personal, for this particular collection.

This work's approach intents to compare content similarity of the blogsites containing Personal (e.g. Personal-inclusive label group), with the collection of blogs that do not contain Personal as one of the pairing categories (e.g. Personal non-inclusive label group). There are a total of 343 distinct labeling categories for the entire collection, of which 241 have shown an association with Personal, including the top categories of Blogging, Finance, Writing and Lifestyle.

To find suitable category replacements for Personal with respect the top four categories, must focus the similarity search among the combination pairs between the top four categories and the remaining 237 categories related to Personal. These pairings, which do naturally occur within this collection thanks to the bloggers collectivism, reside among the Personal non-inclusive label group. There are 11,741 instances of such combinations.

5.2 Vector Space Model for Tag Comparisons

This work adopts the vector-space model in finding a non-`Personal` matching category among `Blogging, Finance, Writing` and `Lifestyle` to replace `Personal` in the blogsite category pairs. Previously had discussed the merits, and limitations, of using vector space for blogsite clustering in Section Blog Clustering's 'Related Work' among various clustering techniques. The task is to take advantage of its capabilities, applying the technique after encoding the blogsite's posts, represented as term frequency vectors that will allow for similarity comparison.

To achieve this feat, the first step in this process requires to decide on the category terms to be encoded. The collection gathering assembled a total of 389,230 distinct terms, which is rather a very large number. Thus, first seek to reduce this bag-of-words through various filtering mechanisms schemes into a manageable size for the tools to be effective.

It should be noted that this collection of terms, as initially collected, were after the use of pre-filtering of typical known stop words for the language, that is, words or terms that are too common for the vocabulary, such as 'and', 'the', and similar others, that contribute no valuable or distinct information.

The filtering of the about a little over 900 known stop words was taken into account by the web-crawler mechanism when making its rounds throughout the web. However, it did not prevent the crawler from capturing other uncharacteristic type terms. These are categorized as those unpredictable or unforeseeable non-words terms, best described as comprising mainly of either a pair or single characters, containing special characters, merged or concatenated terms, mixed numerical, and made-up

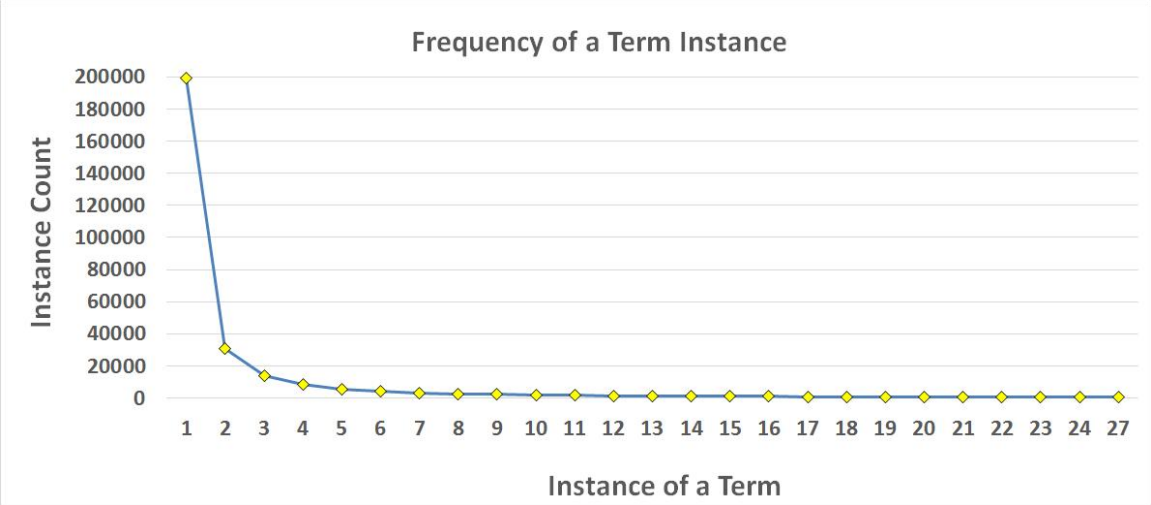

Figure 21. Total instances a word was used once, twice, etc.

words, among these cases. Hence additional filtering techniques were pursued to address the aforementioned cases to help lighten the term collection.

Among the first typical steps would be to perform stemming on the current collection, which would effectively concise the number of terms. Yet, since the collection is still quite considerable, first step is to investigate whether there are terms seldom used within the collection that could be dispensed of, so not to indiscriminately and needlessly invest into stemming such terms.

In this approach, where is meant to identify very infrequently used terms, the scheme is to categorize the smallest term frequency utilization among the different blog postings. The result of this exploration is captured in Figure 21.

In the figure is displayed how often a term might be contained among a set of blog posts. The chart captures how many blog posts within the collection might include a particular term, as indicated by the horizontal axis, and how frequent that case repeats within the entire collection, in the vertical axis.

From this chart, can quickly assert that a rather large number of terms are frequently contained within just a single, or a handful of blog posts. This is most unmistakably noticeable for the single term use case, where this type of event manifests approximately 200,000 times. What this demonstrates is that for a significant number of terms, they are only contained, or rather represented, in a single posting among all blog posts in the collection. That is, there are roughly 200,000 distinct terms, in which each are used only once among all blog postings. Similarly, for the case where a particular term is contained only within two different blog posts, this collection shows that this particular event occurs roughly 30,000 times, or in other words, there are around 30,000 terms that are only contained within a couple of blog posts, and so forth.

This remarkably typifies Zipf's distribution law, which suggests that a term's frequency occurrence is inversely proportional to its rank among all the terms. What this implies, is that typically, the use of a few set of words very frequently, like when describing common or general concepts, akin to the use of stop-words, but mostly tend to use a few choice words only when within a specific context. Hence, the rarity of these words marks them as less than ideal candidates, and as such, though very defining, better of dispensing of them, which can be accomplished through establishing a minimum frequency occurrence threshold setting scenario for the terms.

From Figure 21 can observe that the trend roughly starts leveling in the neighborhood of around 5 postings for a term. As such, instinctively set "5" as the value to filter terms below this threshold. Hence, by adding the initial restriction that any given term should be contained in a minimum of in at least 5 distinct blog posts. This filtering significantly lightens the bag-of-words, which after processing through a Porter stemmer, netted a result of approximately 35,000 terms.

From this set, now it is time to turn the focus to those terms contained in the blog posts that are the confined within the `Personal`-inclusive postings, where represented within all other categories associated for the categories of `Blogging, Finance, Writing` and `Lifestyle`. By using the words exclusively contained within the `Personal` collection, it restrict the search to within the `Personal`-exclusive space with regards to the four targeted categories. This further reduces the stemmed collection to just below 29,000 terms.

Taking into account that the starting collection consisted of 389,230 terms, this new balance represents a substantial reduction in the number of distinct terms. Nevertheless, being that the intent is to use the 'Vector-Space' model (as described in Section 3.4 's conventional approach) to encode the blogs through term frequency vectors, this reduced set of terms is still very sizable and impractical for the vector space model that aims to support.

Reducing this feature set should help simplify the model, thus, increasing the efficiency for the classifiers. The factors to consider when reducing the feature space differ quite apart from the steps incorporated into reducing the size of the term collection. Whereas in the former, stemming and frequency of terms were addressed, in the latter, the focus is on the features that best synthesize/represent a blog. Considering that not all features can be equally descriptive, in which some are more self-evident than others, the next step is into how to select the best features with the greatest capability to best discern between blogs, among members of a class.

For this last task, will rely on Information Gain, a popular text-oriented feature selection technique (Aggarwal and Zhai 2012). It is a metric that can be applied to gauge the discriminatory, discerning power a set of terms within a collection into helping discern among a set of classes, capturing the bits of information per category

prediction on whether or not a term is present, or absent within a document, or blog for this case. It is rooted within the principles of 'entropy' for a data set, as a measure of the level of 'impurity' within the collection.

The IG formula illustrates the components applied into the calculation:

$$I(w) = -\sum_{i=1}^{k} P_i \cdot log(P_i) + F(w) \cdot \sum_{i=1}^{k} p_i(w) \cdot log(p_i(w)) + (1 - F(w)) \cdot \sum_{i=1}^{k} (1 - p_i(w)) \cdot (log(1 - p_i(w)) \quad (5.1)$$

Where:

$I(w)$: Information gain for word w

P_i : global probability of class i

$F(w)$: fraction of the documents containing the word w

$p_i(w)$: probability of class i, given that the document contains the word w

Feature terms in which to calculate the greatest IG value, are those with the greatest discerning power. This work will incorporate information gain as it is readily available in the test environment through WEKA, where the application will help select the top discerning terms.

5.3 Aggregation Strategy Phases

This work will explore alternatives that would help quickly identify the feasibility of the quest into whether it can successfully replace a non-descriptive blog label with more suitable and descriptive label from among categories within the same collection.

Various other techniques have been developed that implement various forms of content-matching, normally at an expensive costs, requiring great computation, and various forms of validation. This technique relies on leveraging on collective wisdom, which should yield more naturally fitting matching content elements that with a lesser computational effort, more closely aligns with the intent of the bloggers for this case. Though the immediate focus is on the `Personal` label, this technique should be applicable to other similar instances in which the intent is to match a label with a more descriptive or other complementing labels.

The nature of the labeled blog environment, which allows a blogger with the option of categorizing his posting with one or more potential labels, transports this environment into the multi-label realm. This type of scenario draws the possibility in which for any given blog posting sample, it could be potentially associated with a subset of label combinations. The number of potential combinations grows dramatically, as the pool of labels grow. The complexity of defining classifiers that would help sort through this label space dramatically increases, as the number of labels grows. This condition gets more aggravated, if many of the categorizing labels are loosely related.

To put in context, the collection of labels for the `Personal` inclusive set, yields a total of 241 labels. Though this number is rather small in comparison to thousands of potential labels in a quick sample from Flickr, or alike, this number of labels would result in approximately a little less than 29,000 possible label-pair combinations to which query a blog posting for a best match. The same number of labels would yield in excess of 2.3 Million label-trios potential combinations, if 3-label combinations were to be explored. As can be seen, this process quickly becomes very inefficient as it escalates, even for a relative small label space, unless some extra help comes to play.

As described by (Bi and Kwok 2013), some efforts have been explored to simplify the label pay load, where label space have been tried to be reduced, simplified, through canonical correlations, principal components, or singular value decomposition, to name a few. Those approaches often are to yield a new reduced, transformed label set that would later prove hard to correlate, understand, or project to the real labels, particularly for this case, with the rigorous hierarchical structure it seeks, need to comply, and should maintain.

Though maintaining this rigorous structure, in the form of the label hierarchy, may seem constraining, it is all to the contrary. The structure provides inside information, as it provide a known relation among the labels that could be exploited for the multi-label classification process. This inside information, which can be deemed, referred as background information, is one that the majority in (Zhang and Zhou 2013) (Sorower 2010) (Zhang and Zhang 2010) (Alaydie, Reddy, and Fotouhi 2012) (Bi and Kwok 2013) (Tsoumakas and Katakis 2007) the multi-label community agree that can facilitate, and boost performance of the classifiers.

It should be noted that as this work dives into these layers of hierarchy, there are limits to the information, as the lines start to blur, as it hones into sub-levels of categorization, making the task of discerning between groups of at the level even more specific, to the point where even parent-child are very closely related. A good example would be when considering an species taxonomy tree. It is relatively easy to set apart cats and dogs ($felidae$ vs. $canidae$ under the $mammals/../carnivorous$ branch), but distinctions starts becoming more difficult as when making incursion into deeper layers, such as beagle, Dalmatian, collie, coyote, wolf, fox, vs. tiger, lynx, cheetah, cougars, Siamese, Persian.

A suggested approach to test the exploits of correlation among labels, is through levels of orders of correlation, in regards to the orders of label relations. The relations addressed are all direct relations, and though some consider constrains in those relations, such as (Alaydie, Reddy, and Fotouhi 2012) hierarchical restrictions, none consider those contributions through collective wisdom, as this process now try to demonstrate. The common technique is to turn each label to be considered, into a series of binary classifiers. This work borrow from these approaches, and loosely apply a simplified version by demonstrating the utility of applying collective wisdom when incorporating into the binary classifiers.

As such, in simplified terms, the overall strategy is to seek identify alternate labels to the less descriptive `Personal` tag for label pairs such as `Personal`-and-`Blogging`, `Personal`-and`Finance`, and alike. To avoid identifying in this search a label whose counterpart would also happen to be `Personal`, this process divides the collection into those whose label is paired to `Personal` (i.e. `Personal`-inclusive) and those who do not (`Personal`-non-inclusive). The process will search solely in this non-inclusive space.

In this process, the label to be complemented (i.e. `Personal`) is defined as the $Target - Label$, and then refer to the pairing label to the $Target - Label$ as the $Anchor - Label$. In this case, `Blogging, Finance, Writing` and `Lifestyle` would become the de-facto $Anchor - Label$(s). Hence, the quest is to find a pairing label to an $Anchor - Label$ to complement the $Target - Label$. This hopeful pairing label would be referred as the $Candidate - Label$.

The process seeks to identify $Candidate - Label$(s), associated to the $Anchor - Label$ while searching within the non-inclusive space. The criteria implemented to identify the $Candidate - Label$ is through content matching, where the content

of the blogs associated to the $Target - Label$-and-$Anchor - Label$ pairings in the $Target - Label$-inclusive space be equivalent, or closely similar, to the contents for the blogs corresponding to the $Anchor - Label$-and-$Candidate - Label$ pairings in the $Target - Label$-non-inclusive space. However, the non-inclusive space is restricted to $Anchor - Label$ associations, as it is from within this group that the processing most likely will find, and should only find, a fitting pairing label, in order to maintain the hierarchical structure.

However, this implies that the contents for the blogs corresponding to the $Anchor - Label$ and associated $Candidate - Labels$ should be sufficiently distinct, and capable to discern with respect to the other label groupings in the non-inclusive space. Otherwise, if this were not the case, neither would be applicable for the inclusive space.

To test the validity of this claim, there are few main considerations needed to be taken into account when performing this quest. First, the search space is within the non-inclusive $Target - Label$ space. Second, and more importantly for this case, the space should be restricted to $Candidate - Labels$ that can be associated to the $Anchor - Label$. The manner in which to implement those association is what the object of this work is all about.

Since the multi-label environment benefits from the background information that can be inferred from the methods described above, through the associations from label relations and hierarchic organizations. It is here it lies where the intent to demonstrate that can obtain equal or better results when incorporating the collective wisdom technique for the label associations to generate the assisting background information than when simply incorporating label relation associations, or hierarchical relations.

This work then is to demonstrate the utility of applying collective wisdom through a series of phases when incorporating into binary classifiers. Then to apply these phases to the TOP-4 categories that had been identified of `Blogging, Finance, Writing` and `Lifestyle`, and the labels that relate to them as per the different phases.

Test Phases:

- PHASE-BASELINE: Only $Anchor - Label$ to $Candidate - Label$ relations considered. This is to represent the baseline.
- PHASE-HIERARCHIC: To only include $Anchor - Label$ to $Candidate - Label$ relations and associations to $Candidate - Label$ as per their hierarchical structure.
- PHASE-COLLECTIVE: To only include $Anchor - Label$ to $Candidate - Label$ relations and associations to $Candidate - Label$ as per collective wisdom analysis.

5.3.1 PHASE-BASELINE

The call of this phase as PHASE-BASELINE, as the initial phase. This will serve as the baseline to which compare the results after the hierarchical and collective wisdom background seeking approaches. In PHASE-BASELINE, do not augment the classifier using background knowledge, that is, the inherent label structure embedded in the data set.

For the BASELINE approach, need only collect $Anchor - Label$ and their direct links to sole relation of $Candidate - Labels$. A sample of the top $Candidate - Labels$ greater than a total of 70 instances is included in Table 7, and also illustrated in the Pareto chart of Figure 22. As a first step to address any label-noise, and spurious labeling concerns, did not include in the collection those label pairs with lower than 20 instances. These conditions had the net effect in which the set only dismissed about

a 10 percent of the total instances, but nevertheless, it represented roughly about a little less than 70 percent of the total label-pairs that had a direct link between the $Anchor - Label$ and the $Candidate - Labels$ for the TOP-4 categories.

From this collection, to further minimize any label-noise concerns, and spurious labeling, segregated and selected the cumulative top 80 percent for each category individually, results of which can be observed on Table 8. These are the label pairs that will be evaluated in the classic $1 - vs - ALL$ fashion for the classifiers.

Notice that from both the Table 7, and also Table 8 and chart of Figure 22, that there is a significant overlap for many of the terms. For example, the $Candidate - Label$ label `Art` is paired to $Anchor - Labels$ `Blogging, Lifestyle` and `Writing`, whereas the $Candidate - Label$ `Business` is paired to all four $Anchor - Labels$. This condition is consistent, and carries into the assessment for the $Candidate - Labels$ corresponding to the ALL component of the $1 - vs - ALL$ collection.

The results described in Section 6 will demonstrate, that these condition should have a difficult time in assessing a good outcome from the classifiers. This should be to no surprise, and very much should be expected, as it will have analogous cases to `Art` and `Business` labels, that will likely affect the training for the condition tested.

5.3.2 PHASE-HIERARCHIC

Referred to this phase as PHASE-HIERARCHIC, as the second test phase in the process. This step will serve as the first test in which to try incorporate background knowledge that should assist in training the classifiers. It also serves to demonstrate, validate and provide support to the argument that adding background information should assist in augmenting the performance of the classifiers.

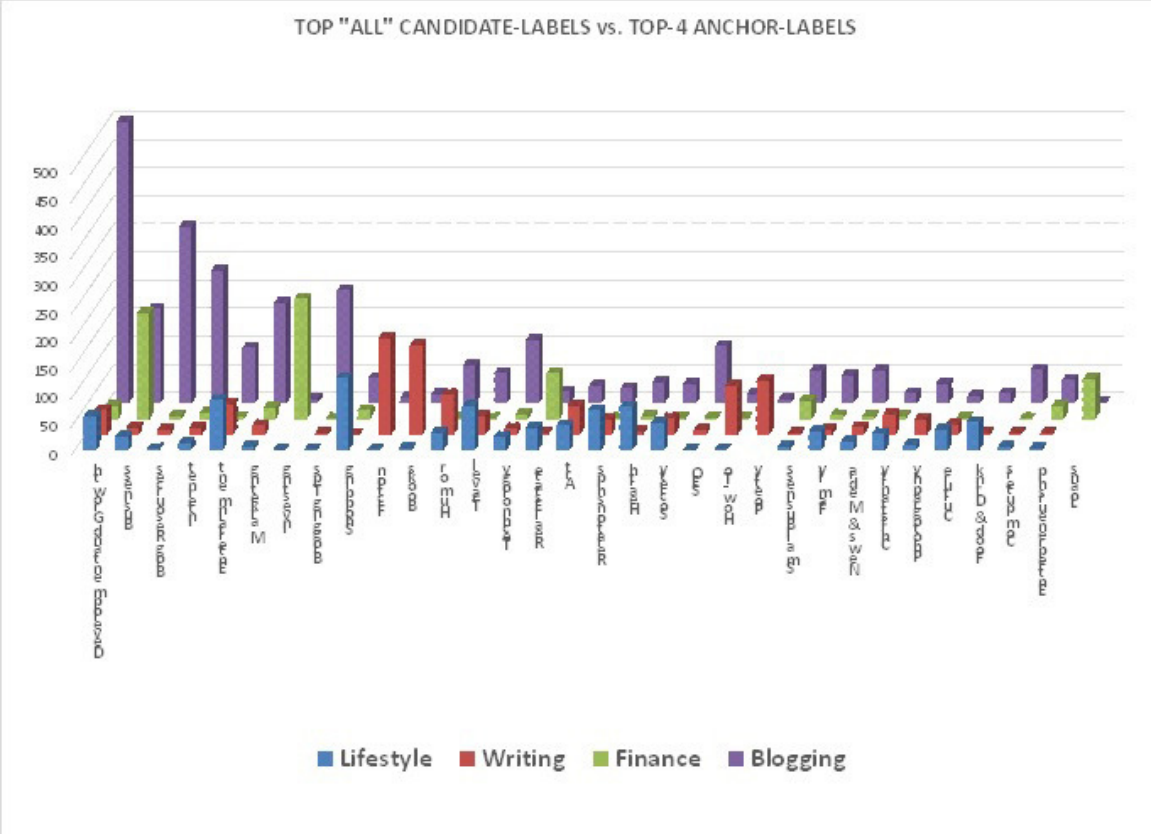

Figure 22. Total Instances of $Candidate - Label$ vs. TOP-4 $Anchor - Labels$

In the BASELINE phase, was only collected $Anchor - Label$ and their direct links to sole relation of $Candidate - Labels$. In the HIERARCHIC-PHASE, contrary to the BASELINE, not only collected $Anchor - Label$ and their direct links to $Candidate - Labels$ relations, it also incorporated the direct links from the $Candidate - Labels$ to any other labels directly associated to the $Candidate - Labels$, and recursively to them as well, that are defined by the links, and hierarchical structure imposed, including layers and sub-levels.

For this phase, also imposed the same label-noise considerations imposed in the HIERARCHIC-PHASE. As such, not included in the collection were those label pairs with lower than 20 instances. This will also result on a dismissal of about a 10

Table 7. *Candidate − Label* vs. TOP-4 *Anchor − Labels*

CANDIDATE-LABEL	Blogging	Finance	Lifestyle	Writing	Total
Development and Growth	1211	24	62	44	1341
Business	168	188	25	13	394
Blog Resources	314	6	3	9	332
Internet	235	12	14	13	274
Entertainment	99	4	90	55	248
Marketing	178	20	8	17	223
Investing	7	214	1	0	222
Blogging Tips	201	1	2	4	208
Shopping	45	16	129	1	191
Fiction	10	0	2	172	184
Books	16	0	5	160	181
Humor	67	1	31	71	170
Travel	54	1	79	33	167
Technology	113	8	24	11	156
Real Estate	21	82	40	4	147
Art	32	0	45	52	129
Relationships	27	1	72	27	127
Health	37	6	78	6	127
Society	35	3	50	29	117
SEO	102	3	1	9	115
How-To	16	2	2	87	107
Poetry	7	0	0	96	103
Small Business	58	33	7	4	102
Family	49	7	34	10	100
News-n-Media	59	5	15	14	93
Christianity	18	5	30	36	89
Photography	35	0	11	28	74
Culture	13	3	38	19	73
Food-n-Drink	18	0	51	4	73
Computers	60	1	7	4	72
Entrepreneurship	42	23	4	3	72
Loans	1	71	0	0	72

percent of the total instances, yet representing roughly about 70 percent of the total label-pairs for the TOP-4 categories.

However, as observed from both Tables 6 and also Tables 7 and 8 and chart of Figure 22, there is a significant overlap for many of the label terms, despite the hierarchical structure imposed by the blogsite. As such, need to add an extra layer to this phase, in which to only associate the $Candidate - Label$, to the $Anchor - Label$ if it has the majority, where is best represented or has the strongest association. The restriction is also applied recursively, where only strong associations to the $Candidate - Label$ are incorporated as well.

The outcome of imposing these conditions are reflected on Table 9. Only the first level tier is included on the table, since the recursive combination of the labels would add complexity for displaying into a table format. These are the label pairs that will be evaluated in the classic $1 - vs - ALL$ fashion for the classifiers for the hierarchical phase.

Note that at first glance at the table immediately suggests that the overlap of $Candidate - Labels$ with respect to $Anchor - Labels$ is mitigated. However, despite the only-the-strongest associations is how a $Candidate - Label$ is associated to an $Anchor - Label$, there exists still a low level interaction of the anchored $Candidate - Label$ with respect to the other non-anchored $Anchor - Labels$. Nevertheless, the results will demonstrate, that the hierarchical enforcement, combined with the strongest association, will greatly improve the outcome from the classifiers for these conditions.

5.3.3 PHASE-COLLECTIVE

For the PHASE-COLLECTIVE, built is a graphical node structure using a graph analysis tool. Through the use of Pajek, which is the same utility incorporated for the

WisColl process. The tool allows to very easily apply, and edit graphs, and visually provides a cue of the state of the graph.

Through Pajek, implemented are various level of graph reduction, displayed as part of the results. Then experiment with different thresholds for the *ALL-label* link strengths[7] range of values. One of the consequences resulting from selecting a link-strength, or threshold, may result in a network re-structuring and reduction depending on the value selected. The re-structuring occurs as a result of removing those links whose line values are below the selected threshold, which can cause a cluster to transform into a smaller cluster and may possibly spawn additional clusters. The cluster size reduction occurs when removing any newly generated isolated nodes, or clusters, that were once linked to at least one other node in the cluster they once subscribed, but belong no more, as their connecting link was removed for falling under the threshold.

As such, as this phase, was started with a grand collection of all the nodes associated with $Anchor - Labels$ and $Candidate - Labels$. Hence illustrated the different sequences, and demonstrated how small clusters start to emerge with each subsequent reduction. Then to select an iteration at a point in which clearly visible clusters associated with the $Anchor - Labels$ are clearly distinguishable, and exhibit minimal interaction between associated $Candidate - Labels$. It is then to identify the $Candidate - Labels$ as the nodes in these sub-clusters that will provide the background knowledge when selecting the blog postings to test/train the classifiers.

[7]number of blog sites that links same pair of labels

Table 8. Top-4 Baseline Phase.

Blogging	Finance	Lifestyle	Writing
Animals	Business	Activism	Academics
Art	Career-n-Jobs	Animals	Art
Beliefs-n-Causes	Cars	Art	Artists
Blog Resources	Commodities-n-Futures	Beauty	Books
Blogging Tips	Corporate	Beliefs-n-Causes	Business
Blogging Tools	Development and Growth	Business	Career-n-Jobs
Business	Economics	Celebrity	Christianity
Computer-n-Video Games	Education and Training	Christianity	Copywriting
Computers	Entrepreneurship	Crafts	Culture
Development and Growth	Global Economics	Culture	Development and Growth
Directories	Investing	Destinations	Education and Training
Education and Training	Loans	Development and Growth	Entertainment
Entertainment	Marketing	Entertainment	Family
Entrepreneurship	Real Estate	Family	Fiction
Family	Shopping	Fashion	Film
Gaming	Small Business	Food-n-Drink	Gay and Lesbian
Health	Stocks-n-Bonds	Gadgets	How-To
Humor		General Health	Humor
Innovation		Green	Independent Publishing
Internet		Health	Internet
Marketing		Home and Garden	Marketing
Moms		Humor	Moms
Music		Interior Design	Music
News-n-Media		Internet	News-n-Media
Photo Blog		Living Well	Non-Fiction
Photography		Mens	Observational Humor
Relationships		Mommy-n-Family	Philosophy
SEO		Moms	Photo Blog
Shopping		Music	Photography
Small Business		News-n-Media	Poems
Social Media		Observational Humor	Poetry
Society		Parenting	Political
Software		Personal Development	Relationships
Sports		Philosophy	Society
Technology		Photography	Spirituality
Travel		Places-n-Geography	Technology
Web Development		Real Estate	Travel
		Relationships	
		Shopping	
		Social Commentary	
		Society	
		Spirituality	
		Sports	
		Style	
		Technology	
		Travel	
		Trends	
		Weddings	
		Womens	

Table 9. Top-4 Hierarchical Phase.

Blogging	Finance	Lifestyle	Writing
Anime	Cars	Beach	Africa
Asia	Commodities-n-Futures	Central-n-South America	Books
Atheism	Corporate	Construction	Cruises
Audio and Video	Economics	Crochet	Fiction
Big History	Emerging Markets	Dance	Graffiti
Blog Resources	Global Economics	Death-n-Dying	How-To
Blogging Tips	Investing	Diet-n-Nutrition	Musicians and Bands
Blogging Tools	Loans	Fashion	Non-Fiction
Car Reviews	Mergers-n-Acquisitions	Fibre	Performance Art
Cartoons	News Only	Fitness	Pets
Classic-n-Collectible	Retirement Planning	Home Improvement	Poems
Cocktails and Spirits	Stocks-n-Bonds	Hyperlocal	Poetry
Computers	Trucking	Interior Design	Primary
Cycling	US Economics	Liberalism	Secondary Education
Development and Growth		Living Well	
Directories		Motor Sports	
eLearning		Natural	
Europe		North America	
Fishing		Outdoor Activity	
Green Politics		Pet Care	
Hockey		Radio	
Hosting		Sailing	
Internet		Soap Making	
Jazz		Style	
Marketing		Toys	
Meme		Weight Loss	
Messianic Judaism			
Online Activism			
Other Languages			
Poker			
Political Humor			
Programming			
Resorts			
Rock			
Scrapbooking			
Scuba Diving			
Search Engines			
SEO			
Sikhism			
Soccer			
Social Media Optimization			
Social Networking			
Templates			
Themes			
Urban Sports			
Veterans			
Web Design			
Web Development			
Web Hosting			

Chapter 6

MULTI-LABEL PROCESS AND RESULTS

Here now described are the steps incurred into validating the process. For each of the mentioned phases, the following steps were performed:

- Identify the target $Anchor - Label$ to evaluate (i.e. `Blogging, Finance`, etc.)
- Create two groups for the classifier, the positive-group (labels associated with the $Anchor - Label$) and the negative group (those not associated)
- Remove low level spurious label associations to identify the top labels within the positive-group
- Identify the top terms associated to the identified top labels to build the positive-group representative feature vectors
- Remove low level spurious label associations to identify the top labels within the negative group
- Remove feature vectors from the negative group whose number of matching feature terms per vector fall below a low-level threshold (i.e. 1, 2, or 3) as the number of feature that contain positive group terms
- Randomize and select two-thirds of the identified positive-group feature vectors
- Randomize and select twice the size of the positive-group of the identified negative-group feature vectors
- Combine the resulting positive/negative vectors and randomize five times.
- Select two-thirds of the combined positive/negative vectors as the 'training' set and train the classifier
- Test the classifier on the one-third balance of the combined randomized vectors.

One other consideration not implicitly mentioned earlier, though was implied, is that the nature of this data set is highly unbalanced. Even the TOP-4 labels only represent about 25 percent of the feature vectors for the `Personal` inclusive group. There are several articles addressing techniques to overcome balance issues. As such, in the steps outlined above, proper measures are taken to address and take balancing into account.

To compensate, best to take a holistic approach. Though one could balance by ensuring that both the positive and negative feature sets are equivalent in size, this would not be too representative of typical, normal encountering cases. Instead, with the understanding that in most typical cases, there are significantly more instances of negative type components, best try to acknowledge that difference, and in a controlled fashion, incorporate slightly more negative samples, than positive ones, to closely resemble real life scenarios, using a 1:2 for the positive-to-negative class ratio.

These adjustments are meant to stay clear, and avoid those cases in which the classifiers yield high accuracy, such as when 90 percent of the features belong to the one same class, deceivingly providing high accuracy marks due to a highly unbalanced classifier.

It is now time to describe process flow in some greater detail. Each $Anchor - Label$ collection consisted in approximately anywhere between 1-to-3K feature postings. For each, it is ensured to have a good representation, of $Candidate - Label$, which was thoroughly "'randomized' a minimum of 5 times.

After randomization, selected 2-Parts for classifier training, and 1-part for testing. For each of the $Anchor - Label$ collection, performed the tests using WEKA into which was implemented a nested filtered classifier incorporating Information Gain, to test Naive Bayes classification for a 'ONE-VS-ALL' type testing in a TF-IDF environment.

After testing, re-combined and re-randomized and repeated the process 5 times, and averaged the results. These results have been summarized in the subsequent tables.

6.1 Results Visualization - Pajek

For the process, systematically experimented with different thresholds for the *ALL-label* link strengths. Each adjustment resulted in a network re-structuring and reduction depending on the link=strength value selected. The re-structuring occurs as a result of removing those links whose line values are below the selected threshold, which can cause a cluster to transform into a smaller cluster and may possibly spawn additional clusters. The cluster size reduction occurs when removing any newly generated isolated nodes, or clusters, that were once linked to at least one other node in the cluster they once subscribed, but belong no more, as their connecting link was removed for falling under the threshold. The clusters were spread and repositioned in the graph in order to align and better illustrate the $Candidate - Labels$ as they became more visible, obvious with respect the $Anchor_Labels$ at hand.

As such, as this phase, was started with a grand collection of all the nodes associated with $Anchor - Labels$ and $Candidate - Labels$. Illustrated are the different sequences, and demonstrate how small clusters start to emerge with each subsequent reduction. Then the process is to select an iteration at a point in which clearly visible clusters associated with the $Anchor - Labels$ are clearly distinguishable, and exhibit minimal interaction between associated $Candidate - Labels$. Thus is then to identify the $Candidate - Labels$ as the nodes in these sub-clusters that will provide the background knowledge when selecting the blog postings to test/train the classifiers.

The following images illustrates the progression of removing nodes whose line values fall below the indicated threshold (i.e. 25, 75, 95,105). As the images progresses, resizing, re-forming, the clusters become evident as the four intended clusters (i.e. `Blogging, Finance, Writing` and `Lifestyle`) emerges. The process is to stop the iterations for reducing string values until it can be noticed when clearly delineated clusters, or if when subsequent reductions, causes an $Anchor - Label$ corresponding cluster to disintegrate, or become childless. Notice in this case, that in the process is better stop when link-strength at 105, which otherwise, for line values greater than 105, greater reduction compromises the clusters density and integrity.

6.2 THE THREE PHASES - Results and Analysis

The summarized/averaged Naive Bayes results using WEKA are displayed next. They are captured per $Anchor - Label$ results, for each of the three phases. It can be noticed from the results, that in all the cases, except one, the HIERARCHIC method is a marked improvement over the baseline.

It can also be noticed from the results, that in every case, except one, the COLLECTIVE method yields the best results. The one instance this is not the case, is in relation to the `Writing` $Anchor - Label$. It can be noticed that for this particular $Anchor - Label$, it is among the smallest clusters per collective-wisdom. Though the `Finance` $Anchor - Label$ seems to be smaller, it has a greater composition with the related $Candidate - Labels$ that shares with `Blogging`. More analysis should be done in this area.

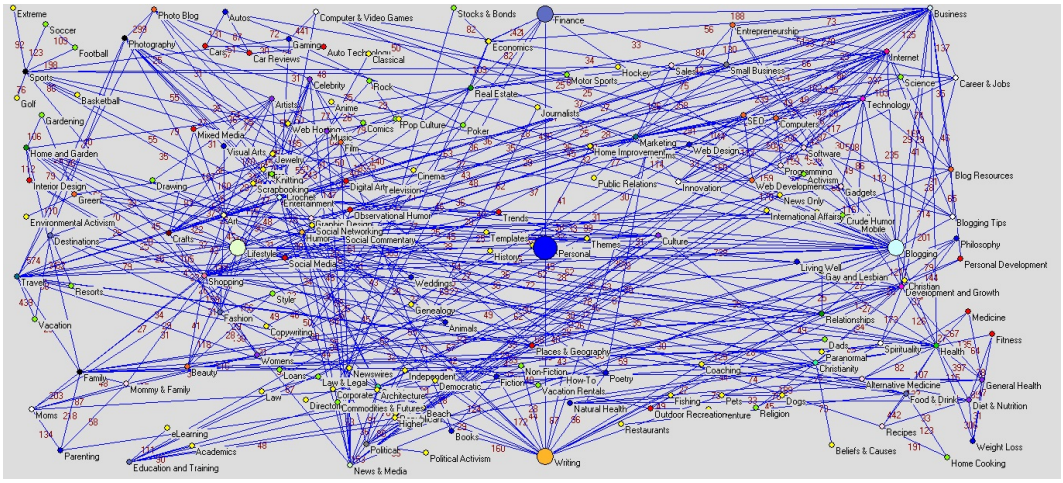

Figure 23. Collective Wisdom Clustering - Nodes with Line Values below 25 removed

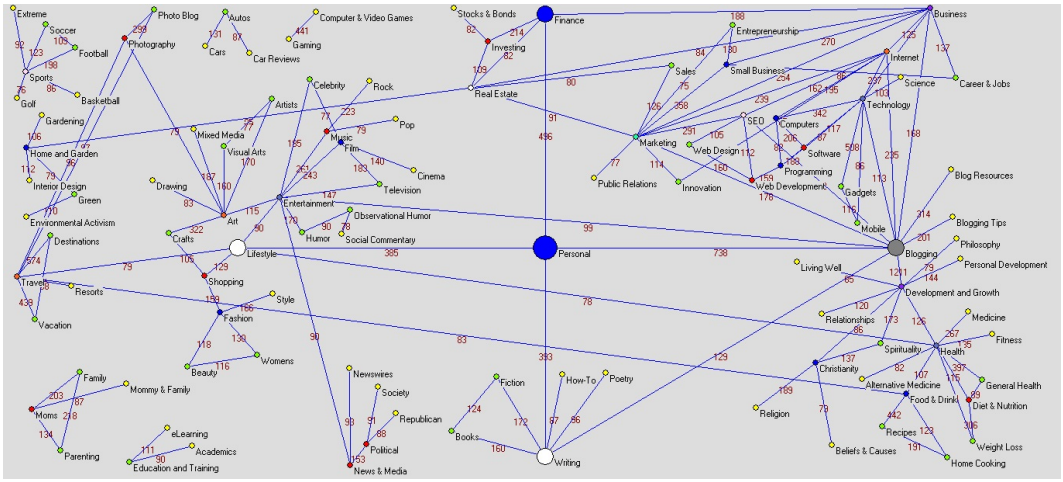

Figure 24. Collective Wisdom Clustering - Nodes with Line Values below 75 removed

Table 10. Blogging Results.

Phase	TP Rate	FP Rate	Precision	Recall	F-Meas.	ROC Area
BASELINE	0.650	0.463	0.643	0.650	0.646	0.640
HIERERCHIC	0.656	0.319	0.694	0.656	0.663	0.732
WISDOM	0.665	0.322	0.685	0.665	0.668	0.733

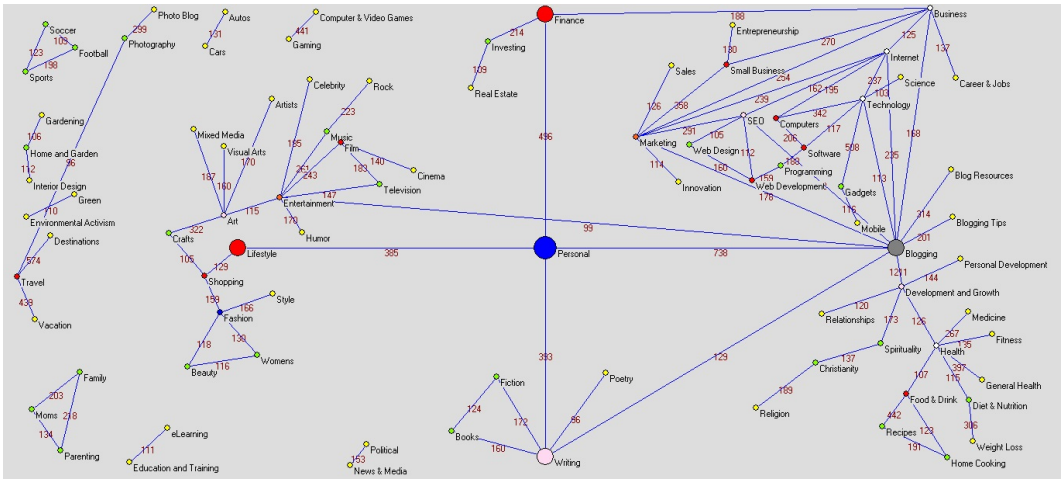

Figure 25. Collective Wisdom Clustering - Nodes with Line Values below 95 removed

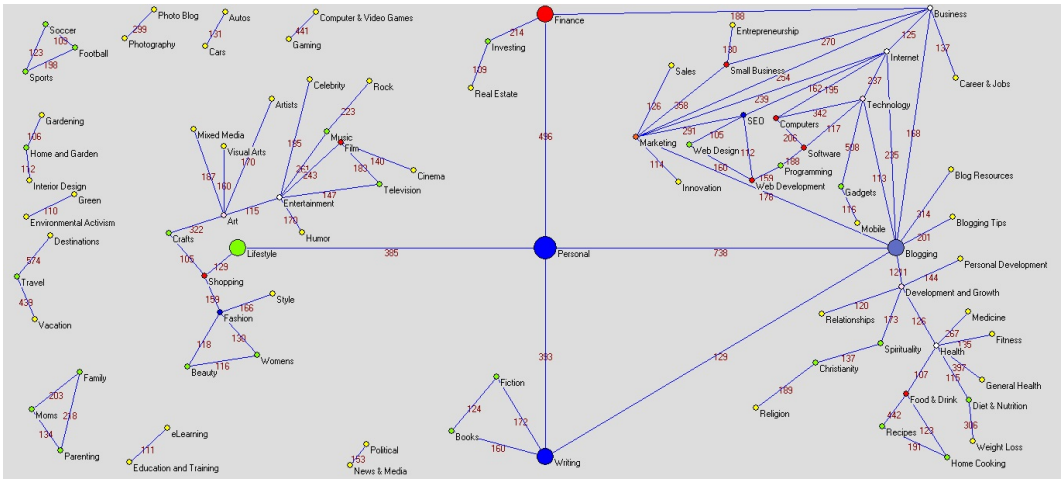

Figure 26. Collective Wisdom Clustering - Nodes with Line Values below 105 removed

Table 11. Finance Results.

Phase	TP Rate	FP Rate	Precision	Recall	F-Meas.	ROC Area
BASELINE	0.785	0.271	0.782	0.785	0.782	0.757
HIERERCHIC	0.772	0.379	0.762	0.772	0.761	0.696
WISDOM	0.815	0.206	0.815	0.815	0.815	0.805

6.3 Extending to Non-Blog Domains

The work and experiments addressed in this document up to this point was developed in a blog exclusive atmosphere, particularly in a data set from BlogCatalog.

Table 12. Lifestyle Results.

Phase	TP Rate	FP Rate	Precision	Recall	F-Meas.	ROC Area
BASELINE	0.662	0.638	0.590	0.662	0.588	0.512
HIERERCHIC	0.688	0.388	0.680	0.688	0.678	0.650
WISDOM	0.737	0.346	0.732	0.737	0.727	0.695

Table 13. Writing Results.

Phase	TP Rate	FP Rate	Precision	Recall	F-Meas.	ROC Area
BASELINE	0.774	0.286	0.774	0.774	0.774	0.838
HIERERCHIC	0.846	0.209	0.846	0.846	0.846	0.912
WISDOM	0.827	0.175	0.828	0.827	0.828	0.900

This section explores the applicability of the reviewed technique upon non-blog type domains.

To support the exploring into these domains, a prescribed requirements would be their ability to entertain a label rich environment intended for sharing, collaboration and categorization as fomented by Web 2.0 in a manner that promotes collective wisdom. Fundamentally, these domains should conform to the primary condition into which a large number of participants independently assign labels to their corresponding Internet entities.

In seeking these environments, some potential non-blog spaces have already been referenced throughout this document. Included in such discussion has been the popular sites del.icio.us and Flicker, which, as has been previously established, were at the forefront of introducing labeling as one of the most prominent staples of Web 2.0. Though labeling is now an almost must-have capability for many web-related entities, this section explores how well other popular non-blog label prone environments conforms to the techniques described.

A significant distinction between these environments, and the results from exploring blogs from BlogCatalog, is that in BlogCatalog, the environment imposes a hierarchy of labels, whereas this is seldom the case for other environments, where

the labeling is rather of free form and very unstructured. Maintaining such structure is a very expensive affair. As discussed, labeling is a personal trait, more so in these environments, in which many labels may have personal meaning just to the author, thus prone to interpretation by others, in a process best described as building 'personomies'.

This freely, unstructured, uncoordinated practice is very much unorganized, unmethodical, and offers no guaranteed commonality. Work from (Wetzker et al. 2010) supports this notion, where in their estimates, they identified that about 20 to 30 percent of users were found to differentially label the same object with little overlap.

However, despite this seemingly form of chaos, in certain instances, with time, some order could actually be attained. The number of terms eventually converge, where the label space could then become finite. An example could very well be as simply of an image posted on Flickr for a nature scene, such as a flower or a sunset, which after all, there are just so many ways into which to describe a flower, the sun, or the clouds, thus creating the label relations that is being sought.

Eventually, all these tagging paves a path to tag clouds and folksonomies. This is a result of these sites extending into patterns in which the popularity of the site is a major factor into their ability to evolve into a compilation of labels, and their capability into developing tag clouds to illustratively depict their tag diversity, their popularity as a reflection of what is trending, and utilization. Per established definition, collective wisdom in this context is defined as the shared, common knowledge arrived at, derived from, by a group of individuals. It is tome now to extend this definition to accommodate folksonomies, and cloud tagging, in the context for non-blog enterprises, a testament to their plurality, as corresponding forms of groups of individuals responsible for generation of the folksonomies and cloud collections. Hence, is possible now to extend

tag clouds and folksonomies as give-away properties that could potentially translate into collective wisdom.

Thus, the task is now to demonstrate the validity of such claims for non-blog sites. The heavy burden for these tests will fall in the likes of Flickr and del.icio.us. Not only has these two sites bare the distinction of being attributed as the forefathers of Web 2.0 labeling, they are also great representative examples of the two forms inherent to folksonomies, best expressed by (Wetzker et al. 2010) distinction into 'narrow' and 'broad' folksonomies, as to mostly distinguish the different tagging rights a site might offer. These tagging rights will ultimately prove to be a key factor into providing diversity, purpose, uniqueness into the label space.

In a 'narrow' folksonomy, only a handful of authorized subjects, such as friends or peers, can incorporate tags into a web entity, whereas in a 'broad' folksonomy, there is very little restriction into who, or what, can be trusted with labeling of the entity. In label property terms, 'narrow' folksonomies promotes greater commonality, at a diversity cost, though not a great penalty when considering more suitable for collective wisdom. Quite the challenge for 'broad' folksonomies, which may provide a lesser common ground, with greater uniqueness and added diversity, as label generated are not within a close community, but rather more general, geared towards the personal preferences of the labeler (a.k.a 'personomies'), which however, the individual labels have been shown to 'converge' on the long run. From the selected two champions, Flickr is a representative of the 'narrow' camp, in which the owner, and a selected group of friend, share and categorize a collection of images, whereas del.icio.us is typical of the 'broader' view camp, in which users individually can bookmark, or label, their preferred sites.

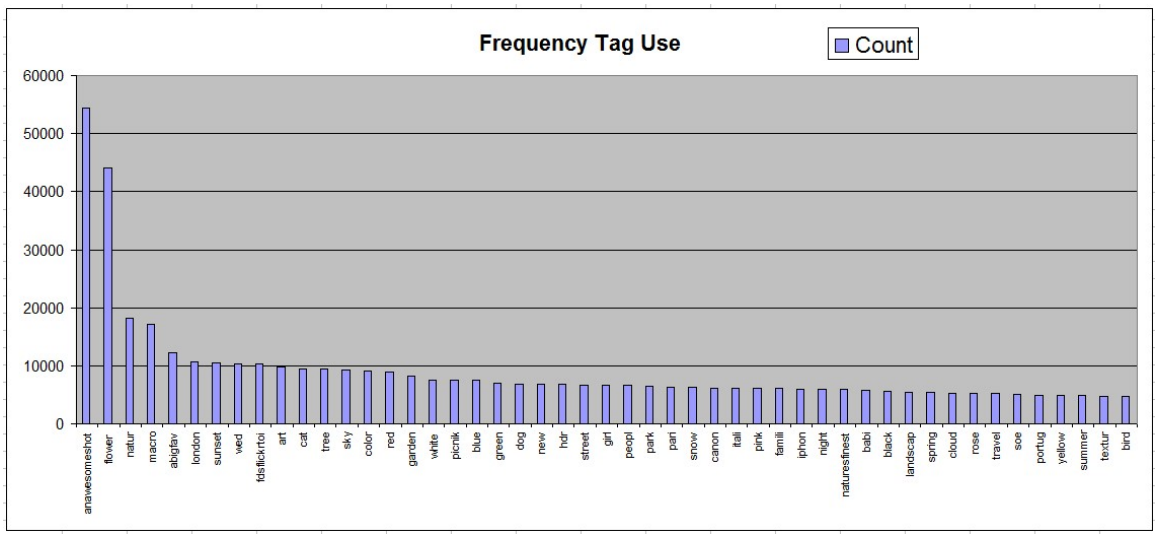

Figure 27. Top 50 Flickr Labels.

The goal now in this exercise is to demonstrate the capability of generating a label-relation graph, as from within a non-blog label space, from which, as has been validated, is the basis for facilitating clustering and label re-labeling through collective wisdom. Per the properties of a 'narrow' folksonomy, this test will incorporate Flickr as the likely candidate to test the non-blog space capabilities of the assertion, as a proof that the technique can be extended into non-blog domains. Shall it were to fail, it greatly reduces the likelihood that a 'broad' folksonomy will be instead successful.

6.4 Flickr as Label Graph

Similarly to BlogCatalog's data collection, crawled the Flickr space of set of users and their friends, collecting the image's URL, their friends, and any labels associated with the images. This collection was significantly noisier than the one collected from BlogCatalog. The number of labels associated to the images ranged to none, to perhaps dozens of them. Herein selected only those images which ranged from 3 labels,

up to a dozen or less, in contrast to BlogCatalog's approach, where only documents with a pair of labels were included. One of the motivators for this approach was the very sparse non-label comments that were sometimes included with an image.

In total, approximately 5.6 million images were captured in this fashion, yielding approximately a little less than 130,000 tags after clean-up of non-text labels. This label space was further reduced to about a 59,000 tags, through stop-words elimination and stemming after translations using Google Translator. Regarding translation, surprisingly, a significant number of images had notations in Spanish, French and Italian, to the tune of approximately 15 percent of the total terms. To address them, since not ready to dismiss them due to the sheer size and potential contribution, first submitted the terms to an English translator using Google's Translate API. Those deemed non English, were then translated to English for the three languages described. Though it was evident other languages were also present, they represented less than 2 percent of the total terms.

Quick analysis revealed that a significant number of images were scenic, or nature related. Hence, labels such as `flower`, `nature` and `sunset` were among the Top 10 labels, as can be seen from the Top 50 Flickr Labels chart. However, upon further inspection of the resulting collection, it was evident from the top terms identified, that a rather large number of users, and contributors, were at awe on the captured image, that were prompt to label it with adjective terms that alluded to its *greatness*, worth as an award winning type image. In all, the number of terms alluding such sentiment, included in the list below, were aggregated into the *anawesomeshot* term label.

- `anawesomeshot`
- `aplusphoto`
- `betterthangood`

- `flickrdiamond`
- `flickrsbest`
- `goldstaraward`
- `photograph`
- `photographi`
- `platinumheartaward`
- `platinumphoto`
- `superbmasterpiec`
- `superfoto`
- `supershot`
- `ultimateshot`
- `vosplusbellesphoto`

Not surprisingly, the *anawesomeshot* term provides little information about the image. This problem sounds quite very familiar, and draws immense parallelism to the `Personal` label problem described in the Multi Label aggregation section. Hence, it would be preferable to replace the *anawesomeshot* label term with a more suitable label replacement. Thus, following those same previous steps, where *anawesomeshot* becomes a $Target - Label$ to replace, it must be first known if a label relation graph could be generated, since dealing in a very different non-blog domain.

Fortunately, the 'Total Nodes Graph post filtering' illustrates the applicability and effectivity of link-strength filtering to the 'Main Nodes Graph'. Thus, if aiming to replace *anawesomeshot* with more suitable labels, next would be to identify the label terms associated with *anawesomeshot*, from which to select the $Anchor - Labels$ and $Candidate - Labels$.

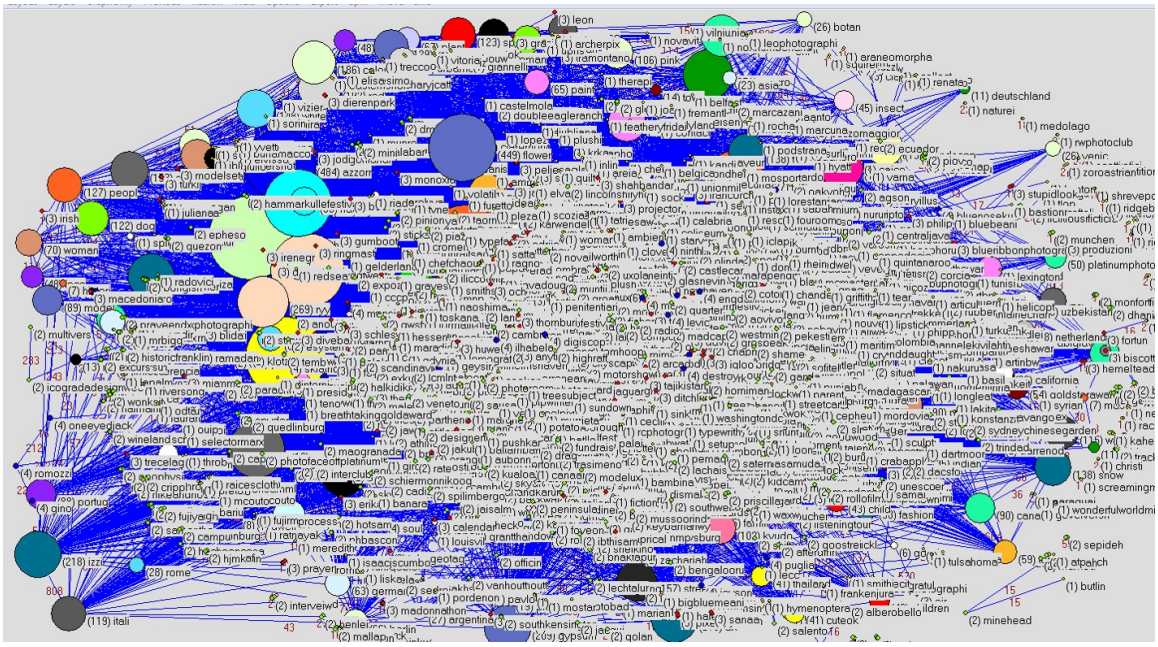

Figure 28. Main Nodes Graph.

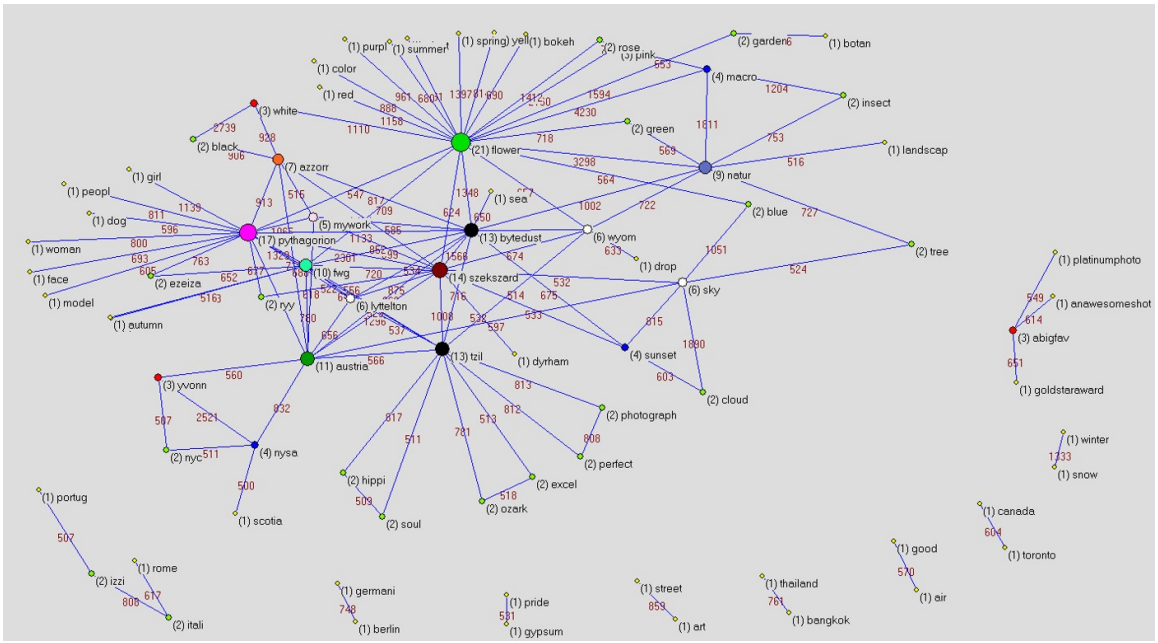

Figure 29. Main Nodes Graph post initial link-strength filtering.

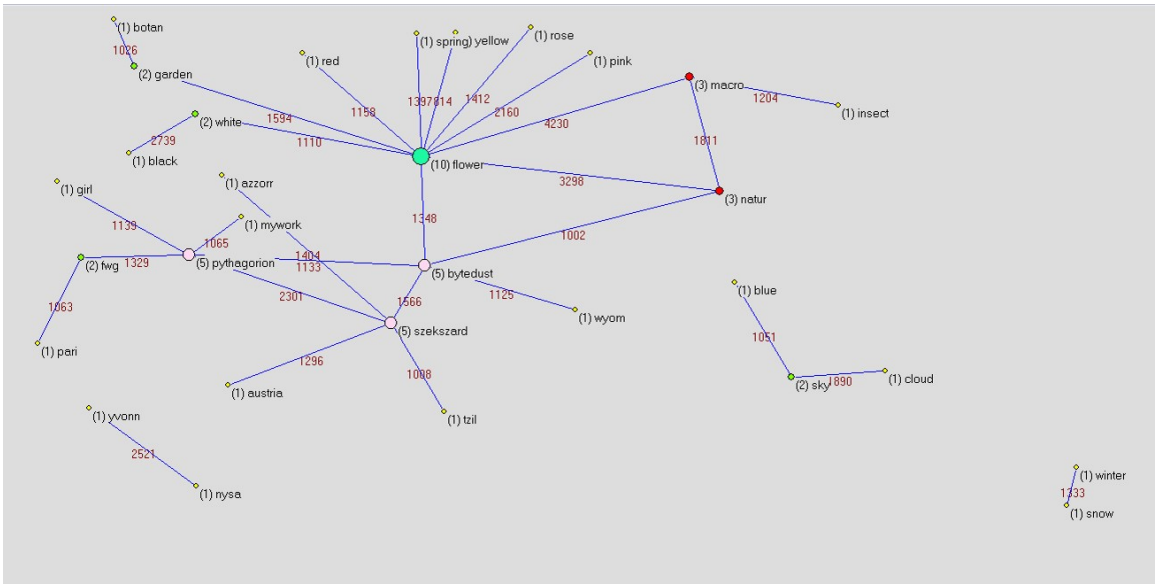

Figure 30. Top $Anchor - Labels$ vs. $Candidate - Labels$.

For this case, the identified top $Anchor - Labels$ are found, and listed below:

- `nature`
- `tzil`
- `flower`
- `bytedust`

The Figure 'Top $Anchor - Labels$ vs. $Candidate - Labels$' illustrates one of the iterations post link-strength filtering identifying potential candidates that could be used to support the $Anchor - Labels$ to $Target - Labels$ to replace with.

Hence, here forth illustrated, and thus demonstrated, the applicability of the technique onto a non-blog environment domain.

Chapter 7

CONCLUSIONS AND LOOKING AHEAD

7.1 Conclusions

In this research, was thus verified the results of label re-assignment by employing collective wisdom. Was then demonstrated that by exploiting the inherent label structure that is generated through the collective wisdom process, can be a source of background knowledge that could assist in achieving a more afine result with regards to the data collection it applies.

It was also investigated and demonstrated the various types of information available in a blog catalog site like user specified tags and labels. Effectively demonstrated the proposal to leverage collective wisdom to generate graphs that represents similarity between labels to create clusters that could help aggregate members of the long tail distribution and share some of the space for some of the influential.

Also demonstrated that could extend this technique to similar $Target - Label$ matching pairs in the collection, such as those corresponding to `Personal`-and-`Humor`, or `Personal`-and-`Entertainment`, and similar ones of interest. Also demonstrated that the technique is not exclussive of blog domains.

Thus below are some of the next steps to consider to further this research.

7.2 Future Work

Future considerations:

- Explore a metric for quality of link-strength (i.e. diversity, density, centrality)
- Define a Pajek/Collective-Wisdom "end-point"
- More aggressive multi-label classifier
- Port technique to other less rigorous multi-label rich environments.
- Explore additional domains, perhaps, not even web related.

REFERENCES

Agarwal et al., Nitin. 2008. *Clustering with Collective Wisdom - A Comparative Study.* Technical report TR-08-004. Arizona State University.

Agarwal, Nitin, Huan Liu, John Salerno, and Philip S. Yu. 2007. "Searching for 'Familiar Strangers' on Blogosphere: Problems and Challenges." In *NSF Symposium on Next-Generation Data Mining and Cyber-enabled Discovery and Innovation (NGDM).*

Aggarwal, Charu C., and ChengXiang Zhai. 2012. "A Survey of Text Classification Algorithms." In *Mining Text Data,* edited by Charu C. Aggarwal and ChengXiang Zhai, 163–222. Springer.

Alaydie, Noor, Chandan K. Reddy, and Farshad Fotouhi. 2012. "Exploiting Label Dependency for Hierarchical Multi-label Classification." In *PAKDD (1),* 7301:294–305. Lecture Notes in Computer Science. Springer.

Anderson, Chris. 2006. *The long tail : why the future of business is selling less of more.* New York : Hyperion.

Andler, Daniel. 2012. "What has collective wisdom to do with wisdom?" *Collective Wisdom: Principles and Mechanisms:* 72–94.

Bansal, N., F. Chiang, N. Koudas, and F. W. Tompa. 2007. "Seeking stable clusters in the Blogosphere." In *Proceedings of VLDB-07.*

Bi, Wei, and James Kwok. 2013. "Efficient Multi-label Classification with Many Labels." In *Proceedings of the 30th International Conference on Machine Learning (ICML-13),* 28:405–413. JMLR Workshop / Conference Proceedings.

Briskin, A., and S. Erickson. 2009. *The Power of Collective Wisdom: And the Trap of Collective Folly.* Berrett-Koehler Publishers. https://books.google.com/books?id=vwx5PAeTxc8C.

Brodley, Carla E., and Mark A. Friedl. 1999. "Identifying Mislabeled Training Data." *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH* 11:131–167.

Brooks, C.H., and N. Montanez. 2006. "Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering." In *Proceedings of the WWW 2006.* Edinburgh, UK: ACM.

Brooks, Christopher H., and Nancy Montanez. 2006. "An Analysis of the Effectiveness of Tagging in Blogs." In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs'06,* 9–14.

Chin, Alvin, and Mark Chignell. 2006. "A Social Hypertext Model for Finding Community in Blogs." In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia,* 11–22. Odense, Denmark: ACM Press. doi:http://doi.acm.org/10.1145/1149941.1149945.

Cutting, D., D. Karger, J. Pedersen, and J. W. Tukey. 1992. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections." In *Proceedings of the SIGIR,* 318 329.

Davis-Stober, Clintin P., David V. Budescu, Jason Dana, and Stephen B. Broomell. 2014. "When is a crowd wise?" *CoRR* abs/1406.7563. http://arxiv.org/abs/1406.7563.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41 (6): 391.

Devaney, M., and A Ram. 1997. "Efficient feature selection in conceptual clustering." In *ICML.*

Dubes, R. C., and A. K. Jain. 1988. *Algorithms for Clustering Data.* Prentice Hall.

Esuli, Andrea, and Fabrizio Sebastiani 0001. 2013. "Improving Text Classification Accuracy by Training Label Cleaning." *ACM Trans. Inf. Syst.* 31 (4): 19. http://dblp.uni-trier.de/db/journals/tois/tois31.html#Esuli013.

Frénay, Benoît, and Michel Verleysen. 2014. "Classification in the Presence of Label Noise: A Survey." *IEEE Trans. Neural Netw. Learning Syst.* 25 (5): 845–869. doi:10.1109/TNNLS.2013.2292894.

Halpin, Harry, Valentin Robu, and Hana Shepherd. 2007. "The Complex Dynamics of Collaborative Tagging." In *Proceedings of the 16th International Conference on World Wide Web,* 211–220. WWW '07. Banff, Alberta, Canada: ACM. doi:10.1145/1242572.1242602.

Hart, Michael, Rob Johnson, and Amanda Stent. 2009. "iTag: a personalized blog tagger." In *Proceedings of the third ACM conference on Recommender systems,* 297–300. RecSys '09. New York, New York, USA: ACM. doi:10.1145/1639714.1639772.

Hayes, Conor, Paolo Avesani, and Sriharsha Veeramachaneni. 2007. "An Analysis of the Use of Tags in a Blog Recommender System." In *Proceedings of the IJCAI.*

Helic, Denis, Christoph Trattner, Markus Strohmaier, and Keith Andrews. 2011. "Are tag clouds useful for navigation? a network-theoretic analysis." *International Journal of Social Computing and Cyber-Physical Systems* 1 (1): 33–55.

Hong, Lu, and Scott E Page. 2012. "Some microfoundations of collective wisdom." *Collective Wisdom: Principles and Mechanisms:* 56–71.

Hotho, Andreas, Robert Jaschke, Christoph Schmitz, and Gerd Stumme. 2006. "Information Retrieval in Folksonomies: Search and Ranking." In *Proceedings of ESWC,* 41126.

Huang, Joshua Zhexue, Michael Ng, and Liping Jing. 2006. *Text Clustering: Algorithms, Semantics and Systems.* PAKDD Tutorial. http://www.ntu.edu.sg/sce/pakdd2006/tutorial/pakdd06-Tutorial%20Text%20Clustering.pdf.

Landemore, H., and J. Elster. 2012. *Collective Wisdom: Principles and Mechanisms.* Cambridge University Press. https://books.google.com/books?id=5vQgAwAAQBAJ.

Lin, Yu-Ru, Hari Sundaram, Yun Chi, Jun Tatemura, and Belle Tseng. 2006. "Discovery of Blog Communities based on Mutual Awareness." In *www'06:3rd annual workshop on webloging ecosystem: aggreation, analysis and dynamics.*

List, Christian. 2012. "Lessons from the Theory of Judgment Aggregation." *Collective Wisdom: Principles and Mechanisms:* 203–229.

MacQueen, J. B. 1967. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the 5th Symposium on Mathematics, Statistics, and Probability,* 281–297.

Mathes, Adam. 2004. "Folksonomies - cooperative classification and communication through shared metadata."

McLachlan, G. J., and K. E. Basford. 1988. *Mixture models. Inference and applications to clustering.* Statistics: Textbooks / Monographs, New York: Dekker.

Mishne, Gilad. 2006. "AutoTag: a collaborative approach to automated tag assignment for weblog posts." In *Proceedings of the 15th international conference on World Wide Web,* 953–954. WWW '06. Edinburgh, Scotland: ACM. doi:10.1145/1135777.1135961.

Mootee, Idris. 2001. *High Intensity Marketing.* SA Press.

O., Zamir, and O. Etzioni. 1998. "Web document clustering: A feasibility demonstration." In *Proceedings of SIGIR.*

O'Reilly, Tim. 2005. *What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software.* http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, September.

Papadopoulos, Symeon, Yiannis Kompatsiaris, and Athena Vakali. 2010. "A graph-based clustering scheme for identifying related tags in folksonomies." In *Data Warehousing and Knowledge Discovery,* 65–76. Springer.

Qamra, Arun, Belle Tseng, and Edward Y. Chang. 2006. "Mining Blog Stories Using Community-Based and Temporal Clustering." In *Proceedings of the CIKM.*

Shi, Lei. 2013. "Trading-off Among Accuracy, Similarity, Diversity, and Long-tail: A Graph-based Recommendation Approach." In *Proceedings of the 7th ACM Conference on Recommender Systems,* 57–64. RecSys '13. Hong Kong, China: ACM. doi:10.1145/2507157.2507165.

Song, Wei, and Soon Cheol Park. 2006. "Genetic Algorithm-based Text Clustering Technique:Automatic Evolution of Clusters with High Efficiency." In *Proceedings of the Seventh International Conference on Web-Age Information Management Workshops (WAIMW).*

Sood, Sanjay C., and Kristian J. Hammond. 2007. "TagAssist: Automatic Tag Suggestion for Blog Posts." In *In International Conference on Weblogs and Social.*

Sorower, Mohammad S. 2010. *A literature survey on algorithms for multi-label learning.* Technical report.

Steinbach, M., G. Karypis, and V. Kumar. 2000. *A Comparison of Document Clustering Techniques.* Technical report TR-00-034. Department of Computer Science and Engineering, University of Minnesota.

Sun, Jiangwen, Feng ying Zhao, Chong-Jun Wang, and Shifu Chen. 2009. "Identifying and Correcting Mislabeled Training Instances." In *FGCN (1),* 244–250. IEEE, January 24. http://dblp.uni-trier.de/db/conf/fgcn/fgcn2007.html#SunZWC07.

Suroweicki, James. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Little, Brown.

Surowiecki, James. 2005. *The Wisdom of Crowds.* Anchor.

Trant, Jennifer. 2009. "Studying social tagging and folksonomy: A review and framework." *Journal of Digital Information (JoDI).*

Trattner, Christoph, Christian Körner, and Denis Helic. 2011. "Enhancing the Navigability of Social Tagging Systems with Tag Taxonomies." In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies,* 18:1–18:8. i-KNOW '11. Graz, Austria: ACM. doi:10.1145/2024288.2024310.

Tseng, Belle L., Junichi Tatemura, and Yi Wu. 2005. "Tomographic: Clustering to Visualize Blog Communities as Mountain Views." In *Proceedings of the World Wide Web.*

Tsoumakas, Grigorios, and Ioannis Katakis. 2007. "Multi-label classification: An overview." *Int J Data Warehousing and Mining* 2007:1–13.

Wetzker, Robert, Carsten Zimmermann, Christian Bauckhage, and Sahin Albayrak. 2010. "I Tag, You Tag: Translating Tags for Advanced User Models." In *Proceedings of the Third ACM International Conference on Web Search and Data Mining,* 71–80. WSDM '10. New York, New York, USA: ACM. doi:10.1145/1718487.1718497.

Xin Li, Philip S. Yu, Bing Liu. 2006. "Mining Community Structure of Named Entities from Web Pages and Blogs." In *Proceedings of AAAI-06.*

Xu, Linli, James Neufeld, Bryce Larson, and Dale Schuurmans. 2004. "Maximum Margin Clustering." In *Proceedings of the Neural Information Processing Systems Conference (NIPS).*

Xu, S., and J. Zhang. 2004. "A parallel hybrid web document clustering algorithm and its performance study." *Journal of Supercomputing* 30:117–131.

Yin, Hongzhi, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. "Challenging the Long Tail Recommendation." *Proc. VLDB Endow.* 5, no. 9 (May): 896–907. doi:10.14778/2311906.2311916.

Yin, Xiaoxin, Jiawei Han, and Philip S. Yu. 2006. "LinkClus: efficient clustering via heterogeneous semantic links." In *VLDB'2006: Proceedings of the 32nd international conference on Very large data bases,* 427–438. VLDB Endowment. doi:http://dx.doi.org/10.1145/304181.304188.

Zhang, M., and Z. Zhou. 2013. "A Review On Multi-Label Learning Algorithms." *Knowledge and Data Engineering, IEEE Transactions on* PP (99): 1.

Zhang, Min-Ling, and Kun Zhang. 2010. "Multi-label Learning by Exploiting Label Dependency." In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 999–1008. KDD '10. Washington, DC, USA: ACM. doi:10.1145/1835804.1835930.