Facets for Discovery and Exploration in Text Collections

Stuart Rose, Ian Roberts, and Nick Cramer

Abstract—Faceted classifications of text collections provide a useful means of partitioning documents into related groups, however traditional approaches of faceting text collections rely on comprehensive analysis of the subject area or annotated general attributes. In this paper we show the application of basic principles for facet analysis to the development of computational methods for facet classification of text collections. Integration with a visual analytics system is described with summaries of user experiences.

Index Terms—Visual analytics, text analytics, faceting theory, keyword extraction, faceted browsers.

INTRODUCTION

Individuals reporting on a changing domain or scenario are faced with the challenge of understanding its dynamics with limited prior information. While they may begin with known features of interest, one of their goals is to ensure that the full set of information relevant to those features is recalled and explored. Text analytics methods that guide and inform users through large information collections and that support dynamic discovery, exploration, and evaluation of relationships within those collections can significantly improve their efforts.

1 BACKGROUND

A facet is generally defined as "*any of the definable aspects that make up a subject or an object"* [1]. In general, facets are used to distinguish objects within a subject area in order to facilitate navigation and exploration. A facet contains an array of nominal values which may be measured for objects within the subject area. These nominal values may also be referred to as foci, bins, or labels. Faceted classification enables the selection and sub-setting of large collections by more than one dimension through the definition of compound subjects. Facet theorists focus on classification of subject areas in order to organize relatively stable knowledge domains and collections. Applications such as faceted browsers that enable exploration of dynamic collections tend to define facets, and their corresponding foci, from objects' attributes.

1.1 Facet Theory

Facet theory developed in the 20th century within the field of Library and Information Science, through the work of S. R. Ranganathan and the Classification Research Group (CRG) [2]. Facet theory focuses on methods of creating faceted classifications as a means of enabling the declaration of compound subjects in order to navigate and explore large information collections. Since information collections are typically organized into taxonomic hierarchies such as the Dewey Decimal System, by allowing the specification of compound subjects faceted classification allows individuals to locate material that relates to subjects otherwise distant in the taxonomy [3].

Ranganathan and the CRG developed separate but related theories for developing faceted classifications. Ranganathan's theory describes three work planes in which a facet classification system is developed. The Idea Plane involves analysis of a subject field into its component parts. The Verbal Plane involves selection of appropriate terminology to express each of these component parts. The Notational Plane involves selection of a notational device to express each of the component parts. The Idea Plane consists of 14 Canons, 13 Postulates, and 22 Principles; the Verbal Plane of 4 Canons; and the Notational Plane of 19 Canons [4].

The CRG modified aspects of Ranganathan's theory, also referred to as "Colon Classification" for its use of colons to separate elements in a compound subject. The CRG's convention for facet analysis is generally understood to provide 7 principles for the choice of facets, 2 principles for citation order, and 3 principles for notation.

Spiteri outlines the differences between the methodologies developed by Ranganathan and the CRG and presents a Simplified Model [5] that combines common elements of the two approaches. The simplified model presents seven principles for the choice of facets, which are summarized in Table 1.

For a subject area or information collection in which the information is changing, such as scientific literature or world news, facets and foci are dynamic and potentially unknown. Engaging in a rigorous exercise to develop a faceted classification would likely yield a classification scheme that is short-lived. Manual facet analysis for a dynamic subject area is unlikely to be suitable for many analysts or the information collections that they access. However the general principles for facet analysis have merit and are likely to provide useful guidance in developing methods of automatic faceted classifications and faceted user interfaces.

In order to support discovery and exploration within dynamic information collections, computational methods of partitioning a collection into coherent groups can yield effective foci for a facet.

2 COMPUTATIONAL FACETING

Referring to the common definition that a facet is a definable aspect of a subject or object, it is possible then to define a facet for any identifiable and measurable aspect of objects within a collection. Such facets and their corresponding foci are easily identifiable when stored as object metadata and accessed through a faceted browser. These may be defined or assigned manually, or automatically determined through computational methods.

A computational method that determines values for a particular aspect of objects in a collection may generate object metadata used as foci for that aspect's corresponding facet. These include methods for determining clusters, groups, themes, keywords, entities, and time foci. Multiple and related methods may populate the same facet (keywords and entities, clusters and groups).

The principles that guide facet analysis enable us to determine the suitability of computational methods for generating a facet and automatically deriving its foci.

An emphasis in facet theory is on the selection of "fundamental categories" and foci that reflect the "nature of the subject". Since faceting dynamic collections focuses on deriving facets from aspects

- *Stuart Rose is with Pacific Northwest National Laboratory (PNNL), E-Mail: stuart.rose@pnnl.gov*
- *Ian Roberts is with Pacific Northwest National Laboratory (PNNL), E-Mail: ian.roberts@pnnl.gov*
- *Nick Cramer is with Pacific Northwest National Laboratory (PNNL), E-Mail: nick.cramer@pnnl.gov*

Table 1. Seven principles of Spiteri's Simplified Model for Facet Analysis	
differentiation	Facets should be clearly distinguished from each other and offer complementary information.
relevance	Facets should directly relate to the intended application and scope of the system.
ascertainability	Facets should be definable and measurable.
permanence	Facets should reflect permanent aspects.
homogeneity	Each facet should represent only one aspect.
mutual exclusivity	Facets should have unique sets of foci.
fundamental categories	Facet foci should be derived based on the nature of the subject. No foci should apply to all objects.

of objects, the nature of the subject for a facet is determined by the computational method and the objects' aspects upon which it operates.

Evaluation of the suitability of a computational method's information for a facet should include an assessment of whether the method's generated information can be realized as foci that reflect and present fundamental categories. Evaluation of the suitability of each foci can then be based on whether each object assigned to the foci fits the "nature of the subject", and whether this is easily discernible for the user.

An emphasis in previous work is that foci be "mutually exclusive". It should be clarified here that mutual exclusivity pertains to conceptual meaning rather than membership, a point that is frequently missed by those developing faceted classifications for objects, in contrast to subject areas. For example a facet for ***large cats*** having the foci ***Panther***, ***Puma***, ***Mountain Lion***, and ***Cougar*** would violate the mutual exclusivity principle because they refer to the same species of large cat. Foci that share common objects do not violate this principle as long as the foci encapsulate separate meanings. For example a facet for ***pet owners*** having the separate foci ***cat owners*** and ***dog owners*** does not violate this principle, because they are conceptually distinct, even though these foci are likely to have members in common.

Foci therefore should have conceptual distance between them. The assessment of conceptual distance is in most cases a qualitative assessment, dependent on the current scope and state of an information collection, as well as the interests, experience, and goals of individual users. For example in some cases an analyst may wish to organize "Raul Castro" and "Fidel Castro" within a single foci, in other cases they may wish to distinguish between them.

2.1 Computing a Themes Facet

In computing a themes facet, our motivation here is to provide more descriptive cues so that users have better insight into the features of the text collection and can explore with greater precision and identify or evaluate more specific relationships. In order to accomplish this, we integrated within IN-SPIRE the algorithm Computation and Analysis of Significant Themes (CAST) [6] which computes a set of themes for a collection of documents based on automatically extracted keyword information.

We integrated the Rapid Automatic Keyword Extraction (RAKE) algorithm [7] to provide this keyword information. RAKE automatically extracts single- and multi-word keywords from individual documents and then provides a set of high-value keywords to CAST which are clustered into themes. Each computed theme comprises a set of highly associated keywords and a set of documents that are highly associated with the theme's keywords.

Whereas many text analysis methods focus on what distinguishes documents, RAKE and CAST focus on what describes documents, ideally characterizing what each document is essentially about. Keywords provide an advantage over other types of signatures as they are readily accessible to a user and can be easily applied to search other information spaces. The value of any particular keyword can be readily evaluated by a user for their particular interests and applied in or adapted to multiple contexts. The top 25 keywords for a collection of over 5000 Voice of America (VOA) News Articles published in the Spring of 2011 are shown in the Figure below.

libyan people, president barack obama, libya, united states, libyan leader moammar gadhafi, gbagbo, ivory coast, ouattara, president obama, forces loyal, government, anti government protesters, pro gadhafi forces, colonel gadhafi, saleh, president saleh, south sudan, shi ite majority, fukushima plant, pakistan, president assad, israel, tokyo electric power company, defense secretary robert gates, libyan capital

Additionally, the top themes for the same collection are listed in Table 2.

Table 2. Top 10 computed themes for VOA News Articles published between February 14 and June 30, 2011.

nDocs	Key Terms
2963	***government***, ***people***, united states, president, country, president barack obama, killed, obama, forces, military, secretary, time
1144	***protesters***, ***security forces***, violence, syria, human rights, protests, saleh, yemen, anti government protesters, president ali abdullah saleh,
901	***libya***, ***gadhafi***, libyan leader moammar gadhafi, nato, rebels, libyan leader, moammar gadhafi, tripoli, civilians, leader moammar gadhafi
697	***pakistan***, ***afghanistan***, al qaida, bin laden, attack, attacks, troops, qaida, pakistani, islamabad
563	***fly zone***, ***security council***, government forces, libyan government, colonel gadhafi, resolution, foreign minister, benghazi, situation,
352	***japan***, ***plant***, radiation, water, earthquake, tsunami, fukushima, nuclear power, disaster
313	***european union***, ***russia***, eu, europe, greece, germany
312	***china***, ***chinese***, beijing, south china sea, asia, vietnam, taiwan, waters, philippines
236	***assad***, ***president bashar al assad***, syrian government, president assad, syrian security forces, damascus, daraa, syrian, reforms, witnesses,
164	***nato forces***, ***afghan government***, afghan forces, afghan president hamid karzai, afghan officials, karzai, afghan people, eastern afghanistan,

3 Faceted Browsers

Facet browser implementations focus on enabling dynamic navigation and exploration through selection of facet foci. While many examples of faceted browsers are available on the web, they tend to focus on static and single collections of data.

An effective faceted browser should facilitate knowledge acquisition about the subject and interconnections of sets of objects within a collection, enabling the user to identify boundaries and bridges between associated objects in a collection. We briefly describe FLAMENCO and Elastic Lists here, as they have been developed as faceting frameworks, and are demonstrated on a common dataset of Nobel Peace Prize winners.

FLexible information Access using MEtadata in Novel Combinations (FLAMENCO) [8] is a search interface framework that exposes category metadata for objects within a collection to a user in order to guide them through exploration of the collection and to organize results of keyword searches. FLAMENCO applies faceted metadata to allow users to refine and expand the current query without interrupting their exploratory flow.

Elastic Lists [9] represents multiple facets as graphical lists of foci within a single faceting container, which animates transitions between user selections. Foci are resorted within facets according to their relative proportion in the current selection. Elastic Lists also includes sparklines to indicate the temporal distribution of objects within foci.

A characteristic of many facet browsers is that facets are located in multiple places across the user interface. Our initial impressions are that this unnecessarily delays a user's assessment of the current selection state as it requires the user to scan and recheck a broad area of the display in order to confirm their selection.

Elastic Lists provides a notable improvement as selections across facets can be read across the faceting container and individual focis' overlap with the current selection are rendered in a single context.

4 Faceting within IN-SPIRE

Based on its ability to enable efficient drill-down we implemented a faceted browser, the Facets Tool, based on the Elastic Lists design. Figure 1 shows the Facets Tool listing a single facet, the top computed themes for the VOA News Articles.

While we considered that merely representing matching counts for foci would lead a user to overestimate connections among large foci, multiple users preferred the simple labeling of matching counts (88) over relative proportions, (88/114). Rendering the baseline histogram with the matching histogram provides a sufficient visual cue of the relative intersection between overlapping foci. For example in Figure 1, the foci for ***asylum seekers*** and ***turkey*** have the same number of matching documents (25) with the search results for ***refugee****, however it is clear that in contrast with the ***turkey*** theme, the majority of documents in the ***asylum seekers*** theme are about refugees.

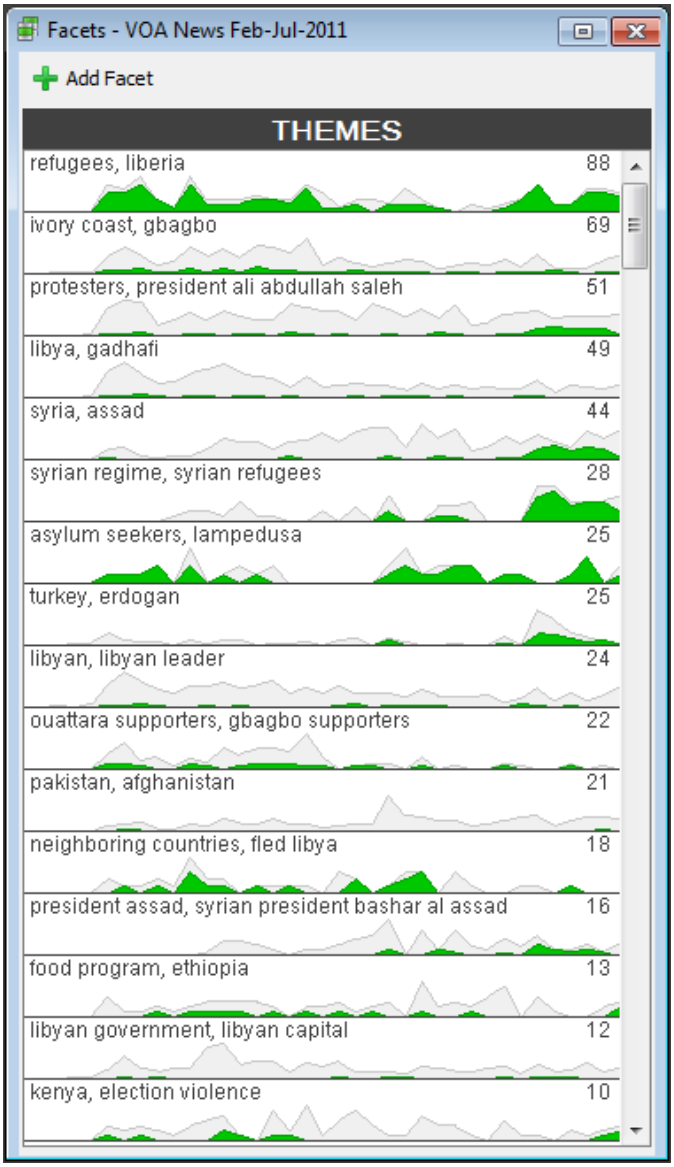

Figure 1. Facets Tool listing top Themes overlapping query results for refugee*.

The sparklines provide further useful temporal information cues to the user. Within the facet for ***syrian regime***, we can see greater overlap in recent months with the search results for ***refugee****.

The temporal distribution of the foci within a facet can be further explored in the Flows View, as shown in Figure 2.

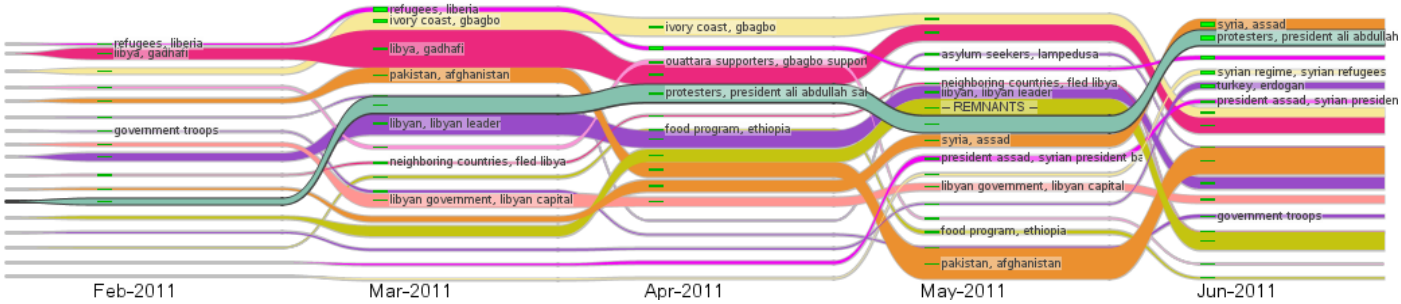

Figure 2. Flows Tool showing temporal distribution of Themes overlapping query results for refugee*.

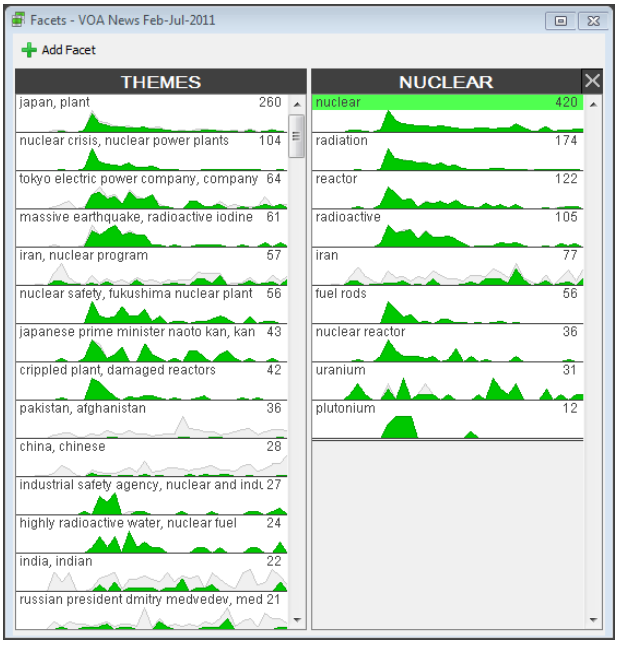

Figure 3. Facets Tool showing facets for Themes and user specified search nodes listed within a Nuclear facet.

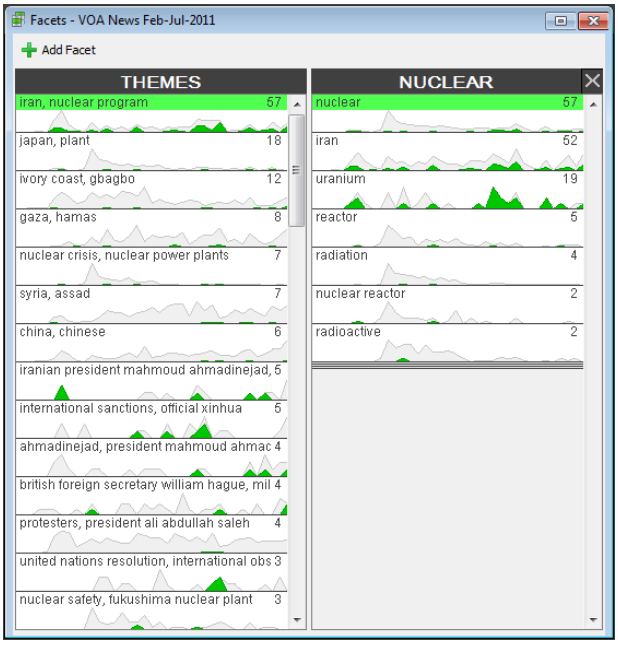

Figure 4. Facets Tool showing facets for Themes and user specified search nodes listed within a Nuclear facet, with "iran, nuclear program" and "nuclear" selected.

The dominant convention for selection in faceted browsers is to perform an intersection of selected foci, regardless of whether selected foci are within the same or different facets. Selection is often referred to as a filtering operation, and objects are filtered out that are not members of each of the selected foci. This convention provides an effective means for navigating to a very specific subset of objects from a larger collection. Its simplicity prevents users from having to specify and remember complex logic about how facet foci are combined.

With multiple facets represented in the Facets Tool, users can effectively work across multiple partitions of the collection. Figure 3 shows the Facets Tool with a Themes facet and a Nuclear facet comprised of search nodes defined by the user. In this case, the user has supplied multiple search nodes as term queries that are applied against the dataset, each producing a set of documents matching that query. Users can then easily select from their defined search nodes and identify which themes are associated with their interests.

In this example we can see multiple themes that overlap with ***nuclear***, and reviewing those themes we can see several similar as well as several distinct themes. Reviewing this list of associated themes, a user can test relationships and choose particular themes to further explore by selecting them in the Facets Tool. Selecting the theme ***iran, nuclear program*** produces a selection that is the intersection of that theme and the search node ***nuclear***, as shown in Figure 4. De-selecting the previous selection will return the Facets Tool to the previous state, enabling a user to explore without significant effort.

5 Conclusion

Faceted browsers provide an efficient and flexible means to understand the dynamics within a collection. Providing users with access to computationally derived facets enables their discovery and exploration of text collections provided those methods take into account basic principles for developing a faceted classification. Additionally, these methods should provide transparency and be easily interpretable by users so that they may more fully leverage the automatically generated facets.

Acknowledgments

The Battelle Memorial Institute manages PNNL for the US Department of Energy under contract DE-AC06-76RL1830. The US government supported this work.

References

[1] Facet. In *Merriam-Webster's online dictionary*. Retrieved from http://www.merriam-webster.com/dictionary/facet

[2] Broughton, Vanda. 2001. Faceted classification as a basis for knowledge organization in a digital environment; the Bliss Bibliographic Classification as a model for vocabulary management and the creation of multi-dimensional knowledge structures. *The New Review of Hypermedia and Multimedia* 2001: 67-102.

[3] David Ellis, Ana Vasconcelos, (1999) "Ranganathan and the Net: using facet analysis to search and organise the World Wide Web", Aslib Proceedings, Vol. 51 Iss: 1, pp.3 – 10

[4] Ranganathan, S. R. (1967). *Prolegomena to Library Classification*. London: Asia Publishing House.

[5] Spiteri, Louise. 1998. A simplified model for facet analysis: Ranganathan 101. *Canadian Journal of Information and Library Science* 23 (1/2) (April-July): 1-30.

[6] Rose SJ, RS Butner, WE Cowley, ML Gregory, and J Walker. 2009. "Describing Story Evolution from Dynamic Information Streams." In IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST) VAST 2009, Oct. 12-13, 2009, Atlantic City, NJ, pp. 99-106. IEEE , Piscataway, NJ. doi:10.1109/VAST.2009.5333437

[7] Rose SJ, DW Engel, NO Cramer, and WE Cowley. 2010. "Automatic Keyword Extraction from Individual Documents." Chapter 1 in Text Mining: Application and Theory, vol. 1, ed. MWBerry, J Kogan, pp. 3-20. John Wiley & Sons, Chichester, United Kingdom.

[8] Elliott, Ame. 2001. Flamenco image browser: using metadata to improve image search during architectural design. In *CHI '01 extended abstracts on Human factors in computing systems* (CHI EA '01). ACM, New York, NY, USA, 69-70. DOI=10.1145/634067.634112

[9] Stefaner, M. and B. Müller, Elastic Lists for Facet Browsers, Proceedings of DEXA '07 18th International Conference on Database and Expert Systems Applications, 2007. FIND07, International Workshop on Dynamic Taxonomies and Faceted Search, Regensburg, Germany. pp. 217-221