*New Review of Hypermedia and Multimedia*
Vol. xx, No. xx, January 2010, 1–12

## RESEARCH ARTICLE

# Capturing the Semiotic Relationship Between Terms

Charlie Hargood, David E. Millard, Mark J. Weal
Learning Societies Lab
School of Electronics & Computer Science
University of Southampton, SO17 1BJ, UK
Email: {cah07r, dem, mjw}@ecs.soton.ac.uk
(*Received 00 Month 200x; final version received 00 Month 200x*)

Tags describing objects on the web are often treated as facts about a resource, whereas it is quite possible that they represent more subjective observations. Existing methods of term expansion expand terms based on dictionary definitions or statistical information on term occurrence. Here we propose the use of a thematic model for term expansion based on semiotic relationships between terms, this has been shown to improve a system's thematic understanding of content and tags and to tease out the more subjective implications of those tags. Such a system relies on a thematic model that must be made by hand. In this article we explore a method to capture a semiotic understanding of particular terms using a rule-based guide to authoring a thematic model. Experimentation shows that it is possible to capture valid definitions that can be used for semiotic term expansion but that the guide itself may not be sufficient to support this on a large scale. We argue that whilst the formation of super definitions will mitigate some of these problems, the development of an authoring support tool may be necessary to solve others.

**Keywords:** Thematics, semiotics, term expansion, narrative, tagging

## 1. Introduction and Background

Folksonomic tagging can provide detailed information about the content of media posted on the web beyond that offered by automatically generated meta data Al-Khalifa and Davis (2006). However the vocabulary used in tagging can often be very specific and it is often necessary to infer what someone means when they search or tag. By expanding what the user is searching for, or expanding the terms used to tag, we broaden the identification process with a range of terms related to what the user might mean and increase our chances of a positive match Buckley (1995).

Web science is about understanding how people use the web so that we might improve it. It is plausible that users imply much more than the literal meaning when they select terms for tags but current term expansion treats terms more as specific facts, making no consideration for their implied meaning or any subtextual use. We believe that by expanding terms on a semiotic basis we acknowledge the themes and concepts beyond the tags literal meaning, and that this can improve a system's understanding of them.

A term can be expanded based on many different properties and relationships and a variety of query expansion projects seek to do this in order to improve the accuracy of searches. Different methods might include synonyms as targets for expansion using thesauri or lexical databases, something investigated critically in Voorhees (1994) using WordNet Miller (1990) and more positively in Buscaldi D. (2005), or statistical methods using words commonly co-occuring with the original

ISSN: 1361-4568 print/ISSN 1740-7842 online

DOI: 10.1080/1361456YYxxxxxxxx
http://www.informaworld.com

term Buckley (1995). Of a variety of methods co-occurrence is often found to be the most successful and this is demonstrated in a review of a variety of approaches in Mandala *et al.* (1999) although it effectiveness is reliant on how the corpus it uses is selected Carpineto *et al.* (2001) and how the co-occurrence frequency is used in expansion Peat and Willett (1991).

In our work we use semiotics as a basis for expanding any given term to other terms that connote or are connoted by it. This semiotic approach expands queries or tags to concepts that might have been implied by the user who created the tag or query.

Take for example the following image[1]:


Figure 1. An example showing how tags might be expanded based on semiotic relationships

The tags used in this image denote some literal signs (such as 'snow' and 'home') but from a semiotic viewpoint these then connote further higher level concepts (they may be describing 'winter' and 'family'). Using an appropriate model it would be possible to expand a set of tags like this to reach a rich network of further implied terms. Treating terms as signs and considering the original term a signifier, and the expanded term a signified is a relationship first discussed in Saussure's original ideas behind the study of semiotics Saussure *et al.* (1966). The idea that a literal denotative sign can be the basis for a connotative sign of something more conceptual was later introduced by Barthes (1957) and provides an explanation behind why one term may be expanded to another through implication.

In our own work we have been exploring how to model themes in narratives, developing a thematic model initially presented in Hargood *et al.* (2008) that represents themes in terms of features, motifs, and themes that are semiotically connected. The idea of structured motifs and themes was originally explored by the formalist Tomashevsky, and although his idea of motifs was much more based around the notions of plot and genre his work Tomashevsky (1965) could be considered direct inspiration for the model. Although collections of knowledge have been used as a basis for expansion before, such as the use of expert ontologies for expansion in Fu *et al.* (2005), our thematic definitions differ from other collections of knowledge (such as expert ontologies) in that they are not attempting to create a canonical objective representation but instead capture subjective connections between terms.

[1]Image and tags taken from www.flickr.com, user findfando

Definitions from expert users who understand the model have already been used with some success in Hargood *et al.* (2009b) to create themed photo montages but in order to get a balanced and broad a set of definitions it would be necessary to have a way for anyone to contribute to a set of definitions. Capturing something so subjective within a formal model is a difficult task, this article reports on the development of a rule-based guide for capturing a thematic definition suitable for use in the semiotic expansion of terms, and an evaluation of the guide through users creating thematic definitions.

## 2. The Thematic Model

The thematic model was initially proposed by our work in Hargood *et al.* (2008). It was developed to provide a thematic underpinning to narrative generation in order to enrich the results of a variety of narrative systems which we explored in Hargood *et al.* (2009a). The model was also further explored in a prototype known as the TMB which was the focus of an experiment in Hargood *et al.* (2009b).

### *2.1 The Model*

The model asserts that a *natom*(narrative atom: a piece of text, image, video, etc.) will contain a number of *features* representing its content. An example would be a photo (the natom) with associated tags (the features). These features then denote the existence of particular *motifs* which in turn, in the context of other motifs, connote the existence of *themes*. Themes may then, along with other themes or motifs, connote further themes. Where as motifs represent devices or generalisations of the tangible features the themes represent higher level concepts (see Figure 2).

There are further rules that govern the semantic quality of definitions; a connotation relationship should not exist unless the contents of all the sub themes and motifs of the connoter are relevant to the connoted. Elements that are connoted or denoted by elements irrelevant to a theme they connote are referred to as 'associated' themes and motifs, these are elements that often co-occur with the theme in question but are not specifically a part of it, and as such should not share a connotation relationship with the theme. This removal of associated elements from definitions of themes helps prevent drift when expanding a given theme. Query drift is a symptom, as noted in Zhou and Huang (2002), of a variety of expansion methods where repeated term expansion through terms falsely considered to be relevant allows the results of the query to become tainted with irrelevant subjects. It is the assessment of potential associated elements and the removal of elements not strictly relevant that stops drift during expansion in the thematic model.

### *2.2 An Example*

Figure 3 shows an example of how a collection of natoms connotes a theme in the terms of the model, in this case a passage of text[1], and two photographs connoting the theme of winter. The features presented are present within the given natoms, it is feasible that the natoms would be tagged with them or that they might be automatically extracted from them. These features literally denote the motifs of snow, cold, and warm clothing. As snow demonstrates many different features might

[1]text from William Shakespears Blow, Blow, Thou Winter Wind

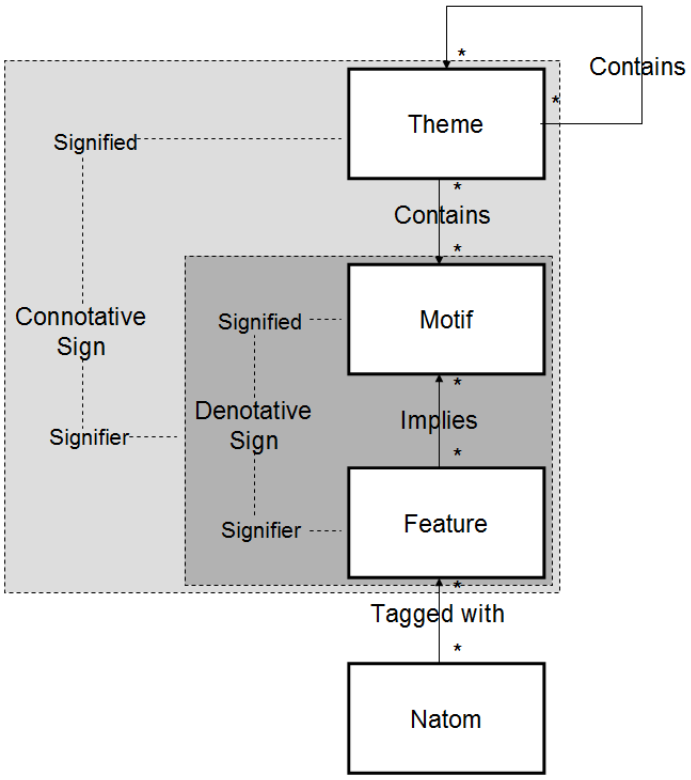


Figure 2. The Thematic Model

denote snow but in this case thematically they serve the same effect. Finally in the context of each other these motifs connote the concept and theme of winter.

### *2.3 The TMB*

The TMB (Thematic Model Builder) is a prototype of a system that uses the thematic model. The TMB compiles photo montages using Flickr [1] based on a title composed of the desired content and one or more desired themes. The resulting montages feature the desired content but are also thematically cohesive. For example 'London' and the theme 'Winter' would produce a montage of London photos that were all wintery. The TMB does this be compiling a large fabula of 30,000 images on the desired content by simply performing a keyword search for it within flickr and then using an instance of the thematic model to select the most relevant images to the theme.

Using instances of the model and its method of calculating thematic quality the TMB successfully performs thematic term expansion through photo montages. However because the semiotic model is static the expansion itself is actually carried out at the authoring stage. As the author builds connotation and denotation relationships between elements in the model they are expanding terms. In connecting a theme to a motif the term associated with that theme is indirectly expanded to all of the features of the motif. Using its models the TMB builds a shopping list of sorts, a list of all features of all relevant motifs to the root theme(s), this list is the thematic expansion of the theme.

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

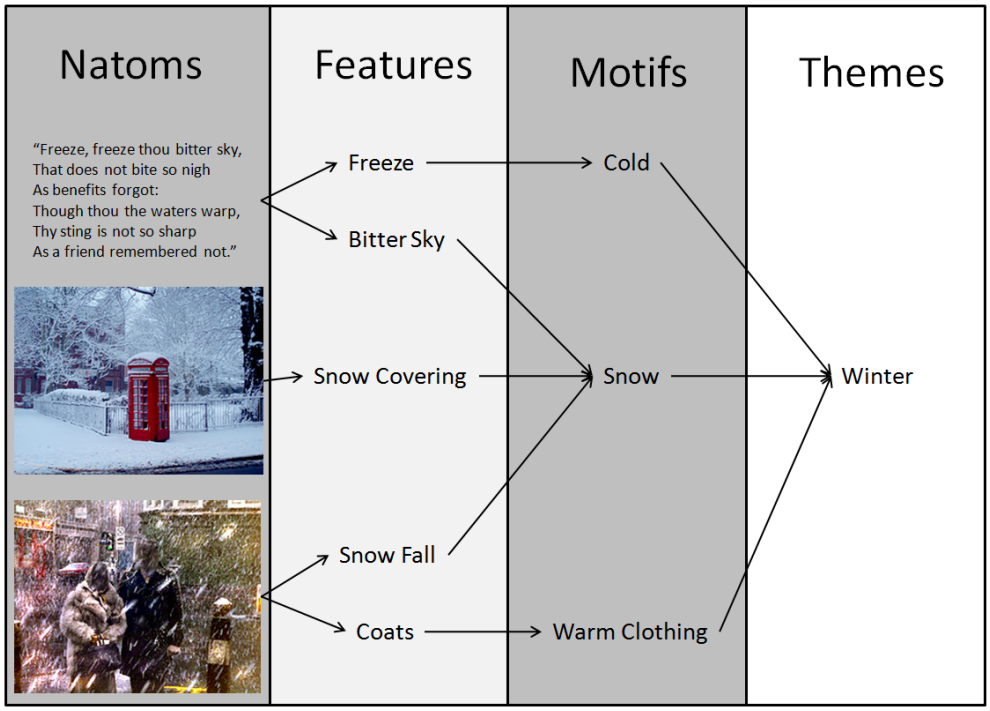


Figure 3. A Worked Example

Based on human-authored models of themes and motifs the thematic expansions avoid several of the negative traits of other methods. Thesaurus based expansion can only expand to terms that have a linguistic connection (synonyms, antonyms, ect.). Co-occurrence isn't limited by such connections but is prone to query drift and over expansion, something that the removal of associated themes in the thematic expansion prevents. The performance of the TMB was evaluated in an experiment reported on in Hargood *et al.* (2009b) and showed promising early results.

## 3. Authoring Experiment

A notable drawback of our semiotic approach to term expansion is that it relies on hand-crafted instances of the model. To this point we have been authoring each theme used in the experiments ourselves according to the rules of the model. A fully working system would need access to a very wide network of themes and motifs, something that is impractical for us to author by hand and so the models would either need to be automatically generated or contributed by a wide community of users. As the thematic model relies on capturing peoples subjective connotations of works it is very difficult to automatically generate such a thing from existing resources on the web, as such the strongest course of action seems to be to develop a way for everyone to contribute to the definitions a system would use.

### *3.1 Creating a guide*

In order for non experts to contribute models to the system they will need a guide or tool that allows them to create definitions based on the rules of the model in an easy step-by-step process. To begin with we elected to develop a guide with a view that anyone could use it to create valid definitions of themes in the terms

of the model. In order to do this first we needed to analyse the process we went through in applying the rules of the model to create a definition. To do this the theme 'danger' was expanded and each decision made in the process of doing this recorded, having completed the process there seemed to be three key tasks:

(1) Identify component elements
(2) Expand sub-themes
(3) Remove associated or weak elements

*Identify component elements*
Most of this task revolves around extracting key parts of the theme in question and classifying each as either a theme or motif based on the semiotic rules of the model. To explain the process in a more accessible way to non experts it was split into two, a word association exercise followed by a classification exercise. This way the contributor could engage in familiar word association to initially expand the theme and then classify the expanded words as themes or motifs based on the rules of the model. A third step was introduced where the contributor grouped together similar elements into a more well defined model.

*Expanding sub-themes and remove associated elements*
These two stages revolve around expanding the sub-themes into further themes and motifs and then removing the weaker associated elements. The process of expansion can be expediated by cutting short the expansion of a sub theme if it becomes apparent that it will later be labeled associated and removed.

This analysis leaves us with a final break down of five stages for defining a given theme in the terms of the model:

(1) *List Associated words*: The contributor spends some time expanding the specific theme into a list of associated words to get a list of related concepts.
(2) *Classify as Themes or Motifs*: The contributor then makes two lists using the results of stage 1 based on the rules of model classifying each as either a theme or a motif.
(3) *Group elements*: The contributor groups together similar elements or those that share a similar purpose.
(4) *Expand Sub-Themes*: The contributor takes remaining theme elements and expands them as they have done the initial theme.
(5) *Remove associated elements*: The contributor removes each theme or motif that is not entirely relevant to the root theme.

Having deconstructed the process the first version of the guide was created, this included an introduction with a short explanation of the model including some specific examples and then a paragraph for each stage explaining what had to be done to complete a definition.

This first version of the guide was then refined through a series of expert reviews. Each review saw a user with experience in modelling use the guide to create a definition of the theme 'danger'. Based on their comments, our observations of the process, and the resulting models changes were made to the guide. Problems included understanding terminology and the elements of the model as well as performing the grouping of elements in stage 3. A series of solutions were attempted with varying success and in the end the final version of the guide was presented more as a table with a worked example along side explaining each stage. The final version was also rewritten in plainer language to tackle the problem of vocabulary and a series of forms were included to guide the forming of the definition.

### *3.2 Methodology*

Having created a guide the next step was to see if a community could use it to collaboratively contribute definitions of themes. To do this we arranged for an experiment where a selection of 15 non-experts (in the sense that they had no experience of modelling abstract concepts) would use the guide to create models for one of five predefined themes. Their definitions would then be analysed to ascertain whether they were valid, and if not what part of the process had been wrongly interpreted and lead to an invalid model. This would give us insight into whether it was possible to capture peoples subjective understanding of themes in a usable form and also whether the guide was sufficient to enable the process.

The test participants were all volunteers from the English department of Southampton University. This meant that they were inexperienced with formal modelling but familiar with themes and thematic relationships. The themes selected to be defined were 'winter', 'spring', 'family', 'celebration', and 'danger'. The first four themes were selected as themes used in the TMB experiments and 'danger' was selected as the theme used in the process of creating the guide.

The participants were invited to attend one of three focus sessions, each of which was approximately an hour long, in which the students created their definitions and were given a very brief introduction. Participants were then given the guide and assigned a theme to define, the themes were distributed from a deck to ensure a random but even allocation. Participants were invited to ask questions but answers were given strictly to clarify the task rather than to influence the decisions made in modelling process. The model definitions were collected and filed for analysis when the participant felt they had finished. The experiment was also passed through the departments ethics committee and granted approval.

### *3.3 Findings*

Having completed the experiment the table in figure 4 summaries the findings, displaying which definitions were valid and in the case of those that weren't which exercises had led to the invalid definitions.

Definitions we're labeled as valid as long as they structurally complied with the models rules, regardless of semantic quality. The notes and forms returned by the participant were analysed for signs as to which stages they struggled on or had questionable results for, the relevant stage to which the participants struggled is also noted on the table.

The results show that just over 50% (8 out of 15) produced valid definitions, however they also show that all the participants except two struggled with the process or produced questionable results for at least one of the stages. Of these stage 4 seems to cause the most problems, followed by stages 2 and 5. It is also to be noted that every participant who produced an invalid definition struggled with stage 4 and this was often the root cause of their invalid definitions.

### *3.4 Analysis*

With these results an analysis of the performance of the guide was possible. Initially we can see that at least half of the participants were able using the guide to produce models that were valid, however almost all participants results showed difficulty with at least one part of the definition process and some of the definitions could be considered semantically poor despite being valid. This shows that while it is indeed possible to capture subjective thematic definitions from people the guide

| Participant | Theme | Valid? | Problem Stages |
|---|---|---|---|
| 1 | Spring | No | 4 |
| 2 | Family | No | 4 |
| 3 | Danger | Yes | 2 |
| 4 | Winter | Yes | 3 |
| 5 | Celebration | No | 4 |
| 6 | Family | Yes | 4, 5 |
| 7 | Spring | No | 4, 5 |
| 8 | Danger | Yes | 4, 5 |
| 9 | Celebration | No | 2, 3, 4, 5 |
| 10 | Winter | No | 2, 4, 5 |
| 11 | Winter | Yes | |
| 12 | Danger | No | 2, 4, 5 |
| 13 | Spring | Yes | 2 |
| 14 | Family | Yes | 2, 3, 4 |
| 15 | Celebration | Yes | |

Figure 4. Summary of experiment results

currently is insufficient to support this process.

Further analysis shows three major observations of the difficulties the participants faced:

- Non-experts struggle with the principles surrounding modelling a concept that are core to this process.
- Participants failed to realise when they had broken a rule of the model presented to them through the guide.
- Some of the heavily subjective decisions necessary in the process are likely to lead to conflicts in definitions and potentially definitions that are valid but of semantically poor quality.

### *3.4.1 Difficulties Modeling a Concept*

Many of the problems faced by participants can be attributed to a lack of experience of modelling a concept as a series of elements and relationships. Participants frequently failed to understand the principle of recursive expansion and that sub themes themselves should also be expanded; this accounts for the large number of participants struggling with stage 4 (which calls for recursive expansion). Definitions, such as the one produced by participant 1 shown in figure 5, often expanded sub-themes on the first layer but not subsequent sub-themes. This led to many models that were invalid simply due to being incomplete as sub-themes were left with no elements connoting them. This can also sometimes lead to definitions including only partially expanded themes as shown in the definition from participant 6 in figure 6. While these models might be valid they are of poor quality as they include themes that are not fully explored.

Participants also struggled to understand child-parent relationships between elements with regards to removing/refactoring associated elements. For example in the definition from participant 6 shown in figure 6 the participant labeled the element ‘emotions’ as irrelevant but did not remove or refactor its parent elements of ‘relationships’ or ‘bond’ (and by extension ‘home’).

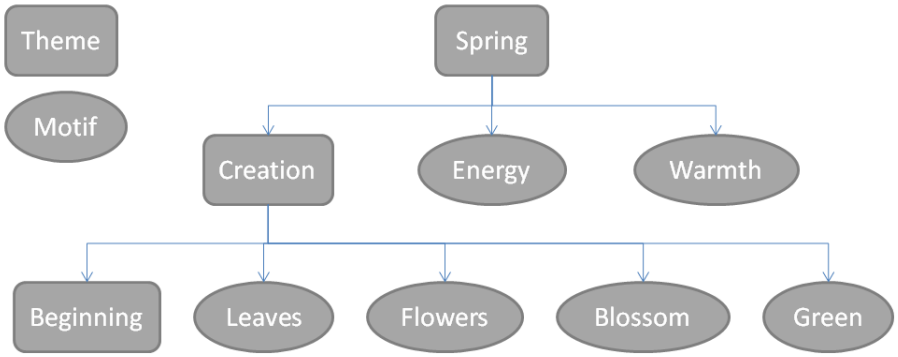


Figure 5. Definition from participant 1

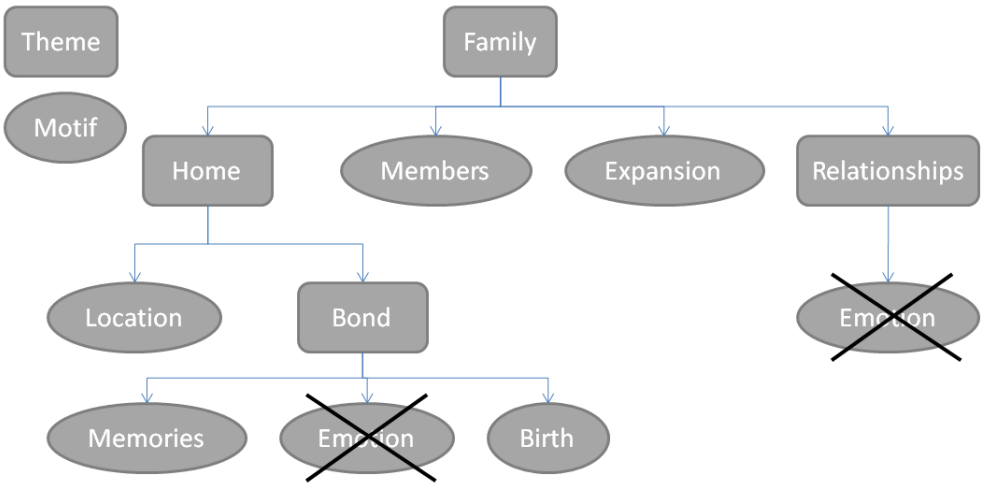


Figure 6. Definition from participant 6

### *3.4.2 Failure to Identify Broken Rules*

A key problem came from a lack of understanding of why a rule was necessary and the ability to recognise when it was broken. In the examples noted previously it is possible that had they been alerted to the breaking of a rule then recursive or incomplete expansion would not be overlooked. There are also knock on effects of these errors that could also be avoided, incomplete expansion often leads to elements not being identified as associated as the irrelevant elements have not been expanded. This is best shown in the definition from participant 12 shown in figure 7 where 'oppressive control' which could quite possibly include elements not relevant to the parent theme is unexpanded, and subsequently left in when it should have been removed.

### *3.4.3 Subjectivity Issues*

We also found further issues arise due to the inherit subjectivity of the process. Much of the trouble participants experienced with stage 2 could be attributed to the difficulty of distinguishing motif and theme. The definition used by participant 12 in figure 7 shows this in that 'death' is classified as a motif despite the fact that stage 2 states that high level concepts should be classified as themes. This could cause conflicts with other definitions that might classify elements differently. The problem of subjectivity also causes difficulties in stage 3 where even with the aid of explicit justifications participants found the grouping of elements difficult,

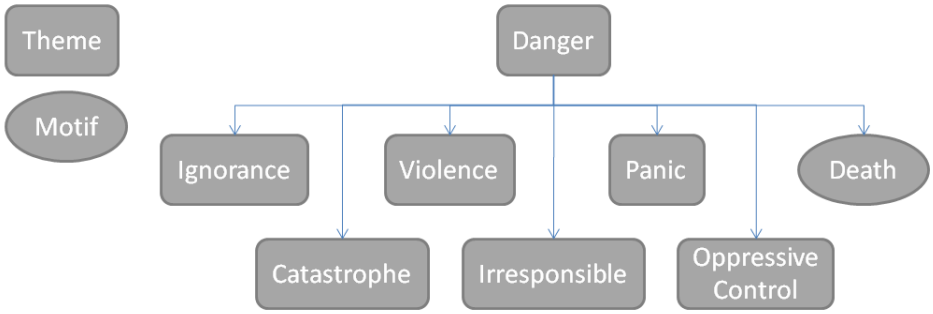


Figure 7. Definition from participant 12

as the role of an element was not always clear. This is demonstrated in figure 8 which shows the definition from participant 4 who has failed to group together hats and scarves despite them both sharing the justification 'worn during'. In other cases elements that might of been grouped together had slightly different subjective justifications and therefore weren't grouped.

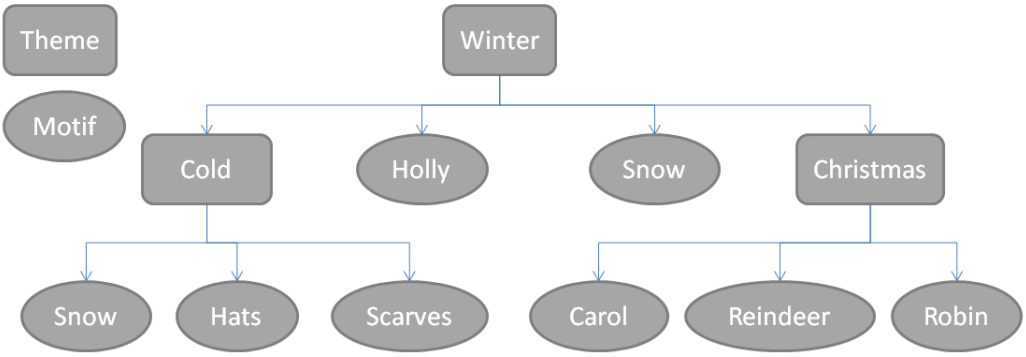


Figure 8. Definition from participant 4

### *3.4.4 Discussion*

The first two observations are related in that they are due to a lack of familiarity with modelling (the level of abstraction required, and the systematic approach needed for completeness). A potential solution to this would be the building of a tool to accompany the guide. This would use a formal schema of the rules of the thematic model to validate definitions, directing users to errors in their creations on the fly. This kind of tool could potentially be applied in many different ways; as well as stand alone authoring tool it could potentially be used alongside tagging systems to allow users to thematically expand tags they used for their media, validating the subsequent definitions.

An authoring tool could help solve the syntactic and structural problems, but not the semantic issues that arise out of differing subjective opinions. However, if large community base of definitions were to be built, a set of super definitions based on the submissions of all participants, then it is possible that such anomalous conflicts could be detected by frequency of occurrence and removed. For example, referring again to the definition of participant 12 in figure 7, if the majority of definitions did not classify 'death' as a motif then such an assertion by a contributor would not be accepted into the super definition. Formation of a super model could also solve the problem of valid models that are only partially expanded such as that of

participant 6 in figure 6 as subsequent definitions could fill out and complete any partial expansions, making invalid definitions partially useful by utilising the parts of them that have been validated. The super model would then form a definition by consensus that could be used by any system utilising the thematic definitions, such as the TMB, and might also allow for weighting of thematic components.

## 4. Conclusions and Future Work

In this work we have explored the possibility of using a semiotic model as a basis for thematic term expansion and to assess whether it was possible to capture the thematic definitions necessary for this to work. By using a semiotic model we gain a greater understanding of the way people are tagging on the web that we can then use to improve systems that use these tags for computation or search.

Previous work has shown that expanding queries on a semiotic basis can improve the thematic relevance of results, however our approach relies on a semiotic model of themes and motifs that is defined by hand, and so a practical method of capturing these definitions from a contributing community is essential. Authoring semiotic models is a very different challenge from the authoring of more factual knowledge representations due to its subjective nature, in particular there are no experts on what different terms connote.

The experiment described in this article suggests that by using the guide developed it is possible to capture people's subjective definitions of themes but also that the guide is often insufficient. While half of the participants produced valid definitions many of them were of a low quality and almost all participants struggled with parts of the process. Much of what the participants found difficult seemed to be with the process of modelling itself, the requirements to think abstractly and to apply rules systematically and exhaustively.

This would suggest that the authoring process is an expert task and that our rules, whilst correct, are insufficient guidance to produce quality definitions on a large scale. The creation of an authoring tool backed up by a formal schema describing the rules of the thematic model could guide authors in applying the rules to their definitions more closely then the guide by itself. The quality of final definitions used by systems could also be improved by the creation of a system that forms definitions submitted by the community into super definitions. This would solve the problem of incomplete definitions by filling them out with the assertions other authors have made as well as improving the semantic quality of the definitions by resolving conflicted definitions by way of popularity. A system that forms super definitions would make even invalid definitions useful, so long as they contained some correctly formed elements and relationships.

The future of this work lies in developing these systems to see if the problems encountered in capturing these subjective definitions using the guide can be mitigated. It also lies in further evaluation of the advantage of semiotic expansion compared other forms of term expansion such as co-occurrence.

Initial experiments with the TMB and keyword search show promising results that semiotic term expansion can lead to a greater understanding of a piece of content. Because of the rules governing the models semiotic expansion demonstrates very little query drift. However this method is still heavily reliant on how its definitions are formed.

In this paper we have shown that while using semiotic term expansion can improve the performance of search there are challenges with creating usable semiotic definitions that are similar to those faced in ontology or taxonomy creation, despite the cognitive differences between subjective thematic models and objective

ontological ones. We have also discussed how these problems might potentially be mitigated through support tools or the community-driven creation of super-definitions. A thematic approach to term expansion affords a system a greater understanding of what users imply when they make a particular query or choose a particular term for a tag, and in our future work we hope to explore how this sub textual understanding can be utilised at a larger scale to help applications find information that is more relevant to their users.

## References

Al-Khalifa, H. and Davis, H., 2006. Folksonomies versus Automatic Keyword Extraction: An Empirical Study. *IADIS International Journal On Computer Science And Information Systems (IJCSIS)*, 1, 132–143.

Barthes, R., 1957. *Mythologies*. Editions du Seuil.

Buckley, C., 1995. Automatic Query Expansion Using SMART : TREC 3. In Proceedings of The third Text REtrieval Conference (TREC-3), 69–80.

Buscaldi D., Rosso P., S.E., A WordNet-based Query Expansion method for Geographical Information Retrieval. In: CLEF 2005 Working Notes Vienna, Austria C. Peters (Ed.), 2005. .

Carpineto, C., de Mori, R., Romano, G. and Bigi, B., 2001. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19 (1), 1–27.

Fu, G., Jones, C.B. and Abdelmoty, A.I., 2005. Ontology-based Spatial Query Expansion in Information Retrieval. In Lecture Notes in Computer Science, Volume 3761, On the Move to Meaningful Internet Systems: ODBASE 2005, 1466–1482.

Hargood, C., Millard, D. and Weal, M., 2008. A Thematic Approach to Emerging Narrative Structure. Web Science at Hypertext08.

Hargood, C., Millard, D. and Weal, M., 2009a. Investigating a thematic approach to narrative generation. DAH at Hypertext 09.

Hargood, C., Millard, D. and Weal, M., 2009b. Using a Thematic Model to Enrich Photo Montages. Proceedings of Hypertext 09.

Mandala, R., M, R., Tokunaga, T. and Tanaka, H., Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion. , 1999. .

Miller, G., 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

Peat, H.J. and Willett, P., 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42, 378–383.

Saussure, F., Bally, C., Sechehaye, A. and Riedlinger, A., 1966. *Course in General Linguistics*. McGraw-Hill.

Tomashevsky, B., 1965. Thematics. *In*: *Russian Formalist Criticism: Four Essays.*, 66–68 University of Nebraska Press.

Voorhees, E.M., 1994. Query expansion using lexical-semantic relations. Dublin, Ireland New York, NY, USA: Springer-Verlag New York, Inc., 61–69.

Zhou, X.S. and Huang, T.S., 2002. Unifying Keywords and Visual Contents in Image Retrieval. *IEEE MultiMedia*, 9 (2), 23–33.