

Advances in classification research, Vol. 17: Proceedings of the 17th ASIS&T SIG/CR Classification Research Workshop (Austin, TX, November 4, 2006), ed. Jonathan Furner and Joseph T. Tennis.

SEARCHING THE LONG TAIL: HIDDEN STRUCTURE IN SOCIAL TAGGING

Emma Tonkin

UKOLN, University of Bath, UK

Abstract

In this paper we explore a method of decomposition of compound tags found in social tagging systems and outline several results, including improvement of search indexes, extraction of semantic information, and benefits to usability. Analysis of tagging habits demonstrates that social tagging systems such as del.icio.us and flickr include both formal metadata, such as geotags, and informally created metadata, such as annotations and descriptions. The majority of tags represent informal metadata; that is, they are not structured according to a formal model, nor do they correspond to a formal ontology.

Statistical exploration of the main tag corpus demonstrates that such searches use only a subset of the available tags; for example, many tags are composed as ad hoc compounds of terms. In order to improve accuracy of searching across the data contained within these tags, a method must be employed to decompose compounds in such a way that there is a high degree of confidence in the result. An approach to decomposition of English-language compounds, designed for use within a small initial sample tagset, is described. Possible decompositions are identified from a generous wordlist, subject to selective lexicon snipping. In order to identify the most likely, a Bayesian classifier is used across term elements. To compensate for the limited sample set, a word classifier is employed and the results classified using a similar method, resulting in a successful classification rate of 88%, and a false negative rate of only 1%.

1. Introduction

Social classification systems employ an inclusive approach to classification. Tagging—the generation and use of free-text metadata for description and discovery of resources—refers to a loose superset of several approaches. In folksonomic tagging, arbitrary text terms are input as metadata describing various objects; the resulting tagging system is used for information resource management, resource discovery and content description.

Certain tagging systems provide a method of employing limited hierarchy in classification. Del.icio.us 'tag bundles' are one such method. Tag bundles, 'tagging of tags', permit a user to bundle a number of tags underneath an umbrella term (eg: a bundle 'photography' might contain 'technique', 'Nikon' and 'club'). This has applications for the process of disambiguation, as well as for neatness; nonetheless, tagging remains a flat namespace. A brief analysis conducted in August 2006 suggests that around 30% of del.icio.us users make at least some use of tag bundling of which just over 10% are prolific (use more than five tag bundles).

Tags are usually represented by a power-law distribution of possible terms, of which a few are extremely popular, but the majority are used only infrequently. Formally, this is an illustration of Zipf's Law (Zipf, 1935), which states that 'in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table'. The 'long tail' of infrequently used terms are sometimes considered to be evidence of a flaw in the concept of tagging, since such terms are unlikely to be retrieved as a direct match for a keyword search. Examples of such tags may result from a lack of synonym control, from spelling errors, from the use of inflected forms and related morphosyntactic phenomena, intentionally specialised terminology chosen according to user convention or consensus, or the result of phrases rendered in the form of a compound term. This paper concentrates on the latter.

Analysis of tagging habits, such as that discussed in (Guy & Tonkin, 2006), demonstrates that social tagging systems such as the well-known sites del.icio.us¹ and flickr² include both formal metadata, such as geotags, and informally created metadata, such as annotations and descriptions. The majority of tags represent informal metadata; that is, they are not structured according to a formal model, nor do they correspond to a formally defined ontology (see Table 1). This approach confers several advantages, such as a low cognitive cost in contribution and minimal resource requirements in terms of infrastructure and management. Data may be combined from several sources due to the heterogeneity of the environment. Due to the informality of the approach, tags may contain incidental information of equal or greater interest to researchers than their intended semantic meaning.

Tag type/%	Words	Simple compounds	Known encodings	Unknown
Flickr	33.8	16	9.7	40.5
Del.icio.us	43.9	23.5	4.3	28.3

1 <http://del.icio.us/>

2 <http://flickr.com/>

Table 1: Tag distribution

On larger collections, 'broad' tagging systems – that is, systems in which many users may tag a given object – quickly attain sufficient terms for simple free-text searches to become effective. This analysis demonstrates a wide span of tagging strategies. In some cases, the choice of strategy may reflect an ad hoc consensus within the community; others reflect reuse of strategies learnt elsewhere. Statistical exploration of the main tag corpus demonstrates that such searches use only a subset of the available tags. For example, many tags are composed as ad hoc compounds of terms.

A particular instance of a compound tag is likely to be unique; that is, the majority of compound terms are applied only once. Many such tags make use of strategies for indicating word boundary (see Table 2), such as the use of a separator character such as a hyphen, underscore or period. However, a significant percentage do not - tags created with separator characters are referred to as 'simple compounds' in Table 1, along with 'short compounds', that is, compounds of two terms. Certain make use of CamelCase - that is, the use of uppercase letters to indicate word boundary. The fact that several tagging systems disallow the use of some or all punctuation marks, and automatically convert tags to lowercase, exacerbates the problem. The subset of these tags that do not make use of a separator of some variety present difficulties to those designing document retrieval systems. A simple keyword search is not well supported, since false positives are likely to result from concatenation – matching across word boundaries – especially with large data sets, where the probability of spurious matches is increased.

Dash	Underscore	Forward slash	Period	Others
39%	25%	14%	14%	8%

Table 2: Common compound separators (del.icio.us)

Certain remedies have been proposed, the majority of which hinge on the principle that the user should be provided with appropriate guidance, such as provision of recommended tagging conventions, error checking and feedback within the interface, and so forth. However, it is possible that optimization of tags for the purpose of keyword search ignores the potential of tagging, for tags could also be seen as a form of digital annotation that responds to a multitude of purposes (Waller, 2003). That compound tags are seen frequently may be taken to imply that the user sees a legitimate purpose for them; one might speculate that a principal use of them is as descriptive metadata, more generally known as free-text annotation.

An illustrative example may be helpful; imagine a photograph of a man holding a dried rose. You decide to call it 'man with dead rose'. The most generic – and thus most likely to receive hits – search terms are therefore 'man', 'dead' and 'rose' – or perhaps 'flower'. Tagging it with these three terms, if the tags are subsequently listed in alphabetical order, provides 'dead man rose'. The dissociation of predicates from the subject of the sentence can lead to startling misapprehensions, for useful contextual information is held in the sentence structure and word order. This relates to existing research on precoordination (Mann, 2000) – that is, the combination of several descriptive terms at the time of indexing, in order to formulate a subject statement or object description – more typically related to other classification methods such as Library of Congress Subject Headings (LCSH). Tags can be linked for search purposes in the manner of a postcoordination, using composite expressions such as tag unions and intersections. However, certain information, particularly in the use of tags as descriptive metadata, is lost in the transition from complex categorisation term to simple search key.

Fortunately, there is no need to characterize this problem as a dichotomy. To quote Mann (2000), 'Neither I nor anyone else is arguing for precoordination rather than postcoordination. We need both browse displays of precoordinated strings and the possibility of postcoordinate combinations of individual elements'. The problem can be approached from another angle, by asking the following question – can the usefulness of little-used tags be improved by analysis?

Analysis of component tags leads to the debate regarding what semantic value, if any, should be placed in the content of any given tag. If this is low, then the value of decomposing a compound term is similarly low. Certain tags are not amenable to this approach, for a variety of reasons; certain tags contain meaning only with respect to a community consensus, whilst others contain information that can be retrieved with no knowledge of the underlying context. This discussion is related to that of the quality of tags in general; in a large collection of tags, in a 'broad' tagging system, community consensus acts to provide a means by which semantics may be inferred from the sum of the tag corpus.

In this paper we describe an approach to decomposition of English-language compounds, designed for use within a small initial sample tag set. Possible decompositions are identified from a generous wordlist, subject to selective lexicon snipping. In order to identify the most likely, a Bayesian classifier is used across term elements. To compensate for the limited sample set, a word classifier is employed and the results classified using a similar method. We demonstrate that this approach, although very simple, is reasonably effective, discuss several shortcomings, and suggest possible improvements. An alternative approach is also identified.

2. The role of preprocessing in keyword indexing

One method of retrieving information from within compound terms involves decomposition of the compound into component terms. These, or an appropriate subset, can then be added to the tags stored for that item. This could be compared to the use of a stemming algorithm on search terms and tags, an approach which for a number of reasons is commonly used in the design of search engines. Firstly, it releases the user from the necessity of searching across each morphological variant of a target term - for example, a search for the term 'watching' matches not only this term but also terms such as 'watched', 'watch' and 'watcher'. Secondly, reducing the number of terms indexed reduces index size and improves efficiency.

This analogy further suggests some probable complications resulting from this approach; stemming algorithms suffer from inevitable limitations related to complexity and exceptions in grammar. Furthermore, to quote Martin Porter (2001), “dictionary-based stemmers require dictionary maintenance, to keep up with an ever-changing language, and this is actually quite a problem. It is not only that a dictionary created to assist stemming nowadays will probably require major updating in a few years time, but also that a dictionary in use for this purpose today may already be several years out of date”. In short, an approach that depends on a preexisting lexicon and formalization of grammar will require maintenance and updating.

One risk specific to this decomposition approach includes the possibility that introduction of incorrectly decomposed tags into a search index would result in an increased level of 'noise'. To offset this risk somewhat, error rate on tag decompositions could be lowered by identifying and discarding uncertain or improbable results. The issue of 'noise' is part of a wider discussion – that of the reliability of user-created tags in general. Given that mechanisms exist for generating or suggesting tags without direct user intervention, and that no formal method of quality control is used in most tagging systems, the mechanism most widely used in 'broad' tagging systems, such as del.icio.us, is that of community consensus. One scenario in which this objection becomes more serious is that in which a tag splitter is used that systematically reproduces the same error, for example, by systematically replacing a common English term by a series of unrelated constituent terms. Assuming the term was common enough, this could seriously affect searches for those constituent terms.

Another potential issue is the homonym problem. This problem is shared by the process of stemming; a

user searching for the unstemmed expression 'watched' does not risk retrieval of irrelevant material on the subject of Rolex or Swatch. However, naïve stemming of their search term strips away contextual information that one might consider as a useful source for disambiguation of the intended sense, and thus of the homonym. Similarly, term splitting with the intention of indexing by component acts to reduce the quantity of information held within the term. As a result, there are implications for index and user interface design, which are briefly discussed later in this paper.

3. Approach

In order to improve accuracy of searching across the data contained within compound tags, a method must be employed to decompose compounds in such a way that there is an acceptable degree of confidence in the result.

Figure 1: Pseudocode for compound decomposition

```
decomposeTerm(String compoundWord)
  find longest prefix of compoundWord in dictionary
  if no prefix matches, return fail.
  record the prefix
  try decomposeTerm(compoundWord except for the prefix)
  if decomposition succeeds, return success.
  else try shorter prefix
```

Possible decompositions are identified by matching across a generous word list (see Figure 1, above), subject to selective lexicon snipping, compiled into an n-ary tree. Each possible decomposition is technically valid, in that it is composed of terms taken from the word list; however, most of the possibilities are meaningless in whole or in part. For example, decomposition of the simple compound 'isaid' provides likely solutions, such as 'i said' or 'is aid', and nonsense solutions such as 'is a id'.

Distinguishing between these candidate splits is the next task. In practice, there are often several valid possibilities. In the majority of cases, it is possible for a human to guess at the most probable of these with fair success and without reference to the object described by the tag. Taking the example of the tag 'photosandsnapshots', where the reader would guess from the first term that the underlying term is 'photos and snapshots', alternative possibilities include 'photo sand snap shots', 'photo sands nap shots' and so forth.

Computationally, some form of decision algorithm is required to perform an analogous task. Rather than looking for a single valid candidate, a class of valid possibilities is defined, and the candidates are sorted into valid or invalid possibilities. Treating this as a document-classification problem has certain advantages, as well as obvious drawbacks; the 'document' is extremely short and contains little useful metadata – the subset of the Unicode set used in the tag, or the date of creation of the tag, may be available depending on the tagging system; the set of possible terms is extraordinarily large.

Initially, the decompositions were classified directly using a naïve Bayesian classifier (see Figure 2), a popular statistical method for classification that is most frequently applied for content classification and filtering purposes such as spam filtering (Sahami et al, 1998). A Bayesian classifier assigns probabilities to features using a supervised learning stage (Witten and Frank, 2005). The majority of candidate decompositions contain improbable interpretations, such as orphaned consonants or unlikely terms; a naïve Bayesian classifier assigns a high probability that statements containing these artifacts are incorrect decompositions.

However, the limited sample set (around 3,000 terms) implied a paucity of training data, too little to support the direct use of this approach. Instead, each decomposition was marked up using a parts-of-speech, or grammatical, tagger – specifically, the TreeTagger (Schmidt 1994). This produces a set of tokens representing the part of speech of each element in the decomposition. Prefixed to each token was the status of that term in the tagger's lexicon; known tokens were prefixed with a 'K', unknown with a 'U'. This meant that the classifier was able to assign separate probabilities to unknown terms acting as a given part of speech, and to known terms. The results of this were then exposed to a suitable decision-making algorithm – the naïve Bayesian classifier. As the output of the tagger contains a limited number of tokens, supervised learning was possible across a smaller set of samples. The filter has no knowledge of word length or of relative placement of words, each of which is potentially relevant and can also be introduced in a similar manner.

Simple ad hoc rules were also successfully used to prune the set of candidate decompositions. For example, average word length is just over 5 characters in a correctly decomposed tag. Tags with a markedly shorter average word length form the bulk of the unacceptable candidates, and can therefore be pruned from the candidate set, or weighted accordingly.

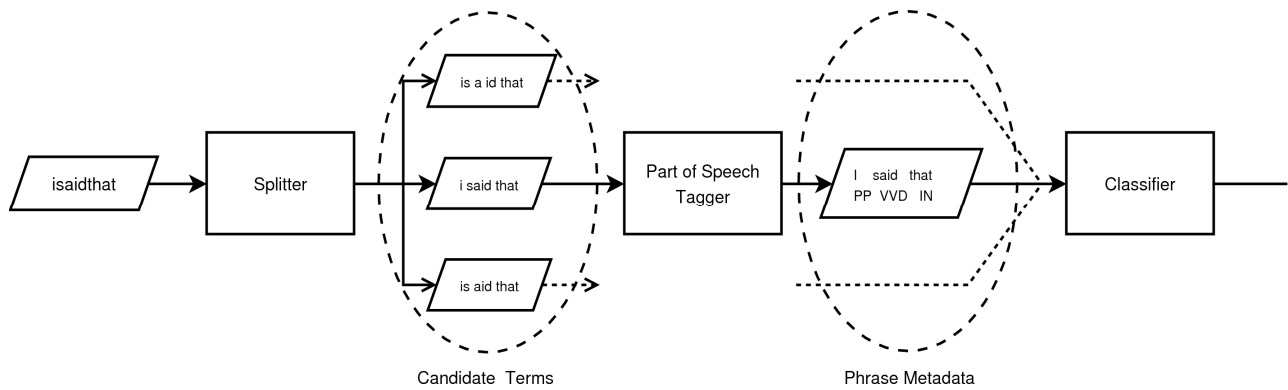


Figure 1: A process for creating and classifying candidate decompositions

4. Evaluation of this approach

The classifier was trained with a small corpus of just under 1000 tags, generated from the decomposition of 30 tags, following the application of the length heuristic. Its accuracy was tested by application to a similar test corpus. 80% of candidates were accurately identified as inappropriate. 8% of tags were accurately identified as likely candidates. 11% of the candidates were wrongly identified as potential candidates – that is, as false positives. However, less than 1% of candidates were wrongly identified as inappropriate – that is, the false negative rate was low.

The false positive rate is of concern, but can be reduced by a number of means. Increasing the threshold value used by the classifier will reduce the number of false positives; on inspection, the classifier assigns extremely high probabilities to likely candidates, so this will not markedly increase the number of false negatives. Taking the output of the classifier as a weighting in parallel with other metrics, such as comparison of average string length, examination of term co-occurrence in a similar corpus, or the use of a Markov chain to calculate the likelihood of a character insertion, will improve the value further.

This approach to tag decomposition has clear limitations. Firstly, the splitter depends on a predefined lexicon of valid terms, meaning that it cannot handle novel terminology or expressions well. This is mitigated slightly by the POS tagger's ability to distinguish between acceptable sentences or sentence fragments that contain a novel expression. However, reliance on this leaves little tolerance for other confounding factors. A more robust solution would require frequent updates to be made to this lexicon. Similarly, the POS tagger itself depends on a lexicon that is far smaller than that used by the splitter, although it classifies unknown terms according to the characteristics implied by its position in the sentence.

Secondly, it favors grammatical sentences as encoded into the POS checker; the tagger infers possible grammatical function of unknown terms, and of known terms used in an unfamiliar sense. This is acceptable in many circumstances, but only a relatively small subset of compound terms are grammatically complete sentences. The majority are sentence fragments or the result of concatenation of associated terms. However, the vast majority are formed in a coherent, if reduced, grammatical structure. Analysis and reuse of this subset grammar would improve the accuracy of the process.

5. Interface considerations

Whilst an optimal level of accuracy is a key goal in development of such a system, in practice the results will contain errors. The quality of the data should be indicated in the interface – when making use of unverified or suspect data, this should be clear to the user. Additionally, the original tag should be made available. If candidate decompositions are displayed to the user, their status should be clearly indicated, particularly an indicator of the level of confidence in the result – confusion or misunderstanding may result from display of incorrectly decomposed tags.

The results of tag decomposition may be revisited at a later time. If users are provided with an appropriate feedback mechanism, corrections may be obtained, such as preference for a specific candidate decomposition. This may indicate that a particular candidate has been mis-classified; however, it may imply only that the incorrect candidate is displayed. Such corrections may be used to revisit the original classifier, amending where necessary. This in turn would result in changes to many other tag decompositions, which additionally carries the advantage that as user-generated feedback is collected, potential changes may be evaluated against existing feedback data and implemented where benefits are clear.

6. Conclusions and future work

We have demonstrated that component terms can be retrieved from certain classes of multiword compound terms, including segmented compound terms and, with reasonable reliability, those which result from simply concatenating words. However, the question that as yet remains is the value of such tags.

For future work, we postulate that it is possible to work without a predefined lexicon and grammatical

structure. We hope to learn a lexicon based on information taken from the subset of tags that use camelCase, underscores or hyphens in order to separate words. These will form the basis of a lexicon and the basis of a grammatical structure. Both grammar and lexicon can be extended as part of the process of decomposing further tags, improving accuracy further and providing some insight into the grammatical structure of compound tagging; in addition to revealing how users tag, such information may bring us closer to understanding the underlying processes. Additionally, such an approach minimises the resource requirements for reuse of the approach across a multilingual corpus.

References

- Guy, M., and Tonkin, E. (2006). Folksonomies: Tidying up tags? *D-LibMagazine* 12(1). Retrieved June 20, 2006 from <http://www.dlib.org/dlib/january06/guy/01guy.html>.
- Mann, T. (2000). Is precoordination unnecessary in LCSH? Are web sites more important to catalog than books? A reference librarian's thoughts on the future of bibliographic control. *Bicentennial Conference on Bibliographic Control for the New Millenium*, Library of Congress. Retrieved August 15, 2006 from http://lcweb.loc.gov/catdir/bibcontrol/mann_paper.html.
- Porter, M. (2001, October). Snowball. Retrieved from <http://snowball.tartarus.org/>.
- Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. *AAAI'98 Workshop on Learning for Text Categorization*.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*.
- Waller, R. (2003, July). Functionality in digital annotation: Imitating and supporting real-world annotation. *Ariadne* 35. Retrieved August 15, 2006, from <http://www.ariadne.ac.uk/issue35/waller/>.
- Witten, I. H. (2005). *Data mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann.
- Zipf, G. K. (1935). *The psychobiology of language: An introduction to dynamic philology*. Boston, MA: Houghton-Mifflin.