Webology, Volume 15, Number 2, December, 2018

Folks Thesauri or Search Thesauri: Why Semantic Search Engines Need Folks Thesauri?

Alireza Noruzi

Ph.D., Editor-in-Chief of Webology, Associate Professor, Department of Knowledge and Information Science, Faculty of Management, University of Tehran, Iran. ORCID. E-mail: noruzi@ut.ac.ir

Abstract

The term ‘folks thesaurus’ was coined as a combination of ‘folks’ and ‘thesaurus’. A folks thesaurus puts terms into context by defining a variety of semantic relationships among the thesaurus terms. The objective of this study is to propose and present a conceptual basis from which it will be possible to build a folks thesaurus. The folks thesaurus takes its terminology and usage from a variety of sources (e.g., users' search queries, traditional thesauri, Wikipedia categories, folksonomies, social tagging, hashtags, and log file analysis of search engines). Folks thesaurus describing subject content can play a vital role in supporting web indexing and information retrieval. Folks thesauri are useful in bridging the gap that exists between the concepts presented by authors and the conceptual terms presented by a user/searcher. Folks thesaurus limits the terms available and increases the possibility that the query will use appropriate terms. If the folks thesaurus has structure in the form of associative or hierarchical tree structure and semantic relationships such as broader terms (BT), narrower terms (NT) or related terms (RT), these may also help the user in navigation through hierarchical semantic relationships and finding an appropriate query expression. If a query is too inclusive, then narrower terms may be substituted to refine the query. On the other hand, a query returning too few results can be broadened through the substitution of broader terms. Related terms may also be an aid in navigation and query construction.

Keywords

Search thesaurus; Collaborative thesaurus; User-generated thesaurus; Social thesaurus; Social tagging; Folksonomies; Social indexing; Semantic relationships

Introduction

The term ‘folks thesaurus’ was coined as a combination of ‘folks’ and ‘thesaurus’. A folks thesaurus puts terms into context by defining a variety of semantic relationships among the thesaurus terms. As with most taxonomies, thesauri define broader and narrower term relationships (hierarchical semantic relationships). In addition, they specify related terms (associative relationships) that allow the user to identify conceptual-semantic relationships among different term groupings. One more type of term relationship in thesauri is the synonym relationship or equivalence relationship, which establishes preferred and non-preferred terms. These three semantic relationships – the hierarchical, associative, and equivalence – work together to enrich a folks thesaurus and make it far more than a simple wordlist.

A folks thesaurus is polyhierarchy; the system is much more flexible but complexity increases and ambiguities have to be solved. Completely cultivated from and nourished by the collective intelligence of users (like Google Translate), a folks thesaurus is a bottom-up approach that is markedly different from traditional thesauri. Constructing a folks thesaurus, however, is a difficult process which can be facilitated through the application of users' search queries (the past search activities, i.e., previously submitted search keywords of users), Wikipedia categories/contents, folksonomies, social tagging, hashtags, traditional thesauri, and log file analysis of search engines, ontologically organized so that the a priori relationships between concepts (terms, keywords and tags) are made explicit, to be used in information retrieval and recommender systems, providing useful access points.

A folks thesaurus is a subset of the lexicon of a natural language, containing a store of words of preferred and non-preferred terms produced by the process of vocabulary control. A preferred term (also known as subject term, main term, or index term) is a term used consistently to represent and designate a concept. One of two or more synonyms or lexical variants of a term may be selected as a preferred term. Preferred terms are linked together through hierarchical semantic relationships (broader terms and narrower terms), associative relationships (related terms) and inter-language equivalence relationships (e.g., interlanguage links from a search query on Google in one *Wikipedia* language to an equivalent entry in another language, for example, French and English equivalents of the term: bibliothécaire / librarian). In other words, a folks thesaurus is a collection of selected vocabulary (preferred terms or descriptors) by the folks (users) for the folks, with links among broader, narrower, synonymous, equivalent, and other related terms. It is a semantic tool that can be used for web indexing, web information retrieval, and search query expansion. It is basically a selection of the basic vocabulary in a field supplemented with information about synonyms, homonyms, generic terms, part/whole

terms, associative terms and other information (e.g., frequency of terms, and frequency of search hits on a given search engine such as Google).

Terms selected should have user warrant in that they are used frequently by users. The user warrant systems are based on the terms that are of interest to the user. User warrant is the guiding principle for the selection of the preferred form of a term. User warrant is the inclusion of a vocabulary term in a folks thesaurus based on use by folks (users). Each term or descriptor included in a folks thesaurus should represent a single concept or unit of thought. Thus, the folks thesaurus incorporates terms and usages from the folks terms (Noruzi, 2007).

Terms in a folks thesaurus are selected from user keywords or tags as well as interviews, print and electronic literature. As these information sources are heavily informed by user warrant, there is often little difference between the terms most commonly used in the texts and those employed by the folks (users). For this reason, literary warrant is the guiding principle for the selection of the preferred terms and user warrant should be consulted to ensure the terms reflected common usage and to determine non-preferred terms. The scope of terms or descriptors should be restricted to selected meanings within the domain of the thesaurus. In the case of ambiguously defined terms or multiple terms for the same concept, an online dictionary (Wiktionary or Wikipedia) should be used to establish the preferred term.

The objective of this study is to propose and present a conceptual and practical basis from which it will be possible to build a folks thesaurus.

Term Relationships in a Folks Thesaurus

A folks thesaurus is a controlled vocabulary of folks terms in natural language. In other words, a folks thesaurus is a controlled vocabulary arranged in a known order and structured so that hierarchical, equivalence, associative, and homographic relationships among terms are displayed clearly. The primary purposes of a folks thesaurus are: (a) to facilitate web information retrieval; and (b) to achieve consistency in the web indexing. A folks thesaurus rules are:

1. Use a limited list of preferred terms, but plenty of entry terms:
   link these with *USE* and *USE FOR* (*UF*) relationships.
2. Structure terms of the same type into semantic hierarchies:
   link these with *Broader Term/Narrower Term* (BT/NT) relationships.
3. Remind users of other related terms to consider:
   link these with *Related Term/Related Term* (RT/RT) relationships.

Like a wordlist, a folks thesaurus is a tool based on the users terminology which help and recommend the user to choose terms to enter into the search box of a search engine. However, unlike a wordlist, a folks thesaurus allows terms, related by a similar subject, to be grouped together into semantic hierarchies and cross-referenced to other groups of related terms which may be relevant to the subject (i.e., search queries). A folks thesaurus provides the user with a single preferred term to use where there is a choice of several terms with the same or similar meaning. Through the use of semantic hierarchies, allows terms to be searched and browsed at a general or specific level, depending on the level of thesaurus terms; and the user will be directed to the preferred term when selecting and searching non-preferred terms. A folks thesaurus is a dynamic tool, which can be developed and updated by the addition, amendment and deletion of terms, relationships or hierarchies as dictated by users.

The traditional thesaurus is the user's guide to the terminology used by subject specialists to describe the content of the database as well as the discipline the user are studying. Some databases are indexed by a thesaurus of descriptors, or index terms, which have been carefully chosen and controlled by the information producers of the database (e.g., ERIC, Embase, Ovid, and ProQuest). But folks thesaurus is the terminology used by web users.

As the control vocabulary, the folks thesaurus standardizes the terminology used for searching. A folks thesaurus typically has a hierarchical tree structure in which general topics branch off to terms with narrower meanings. For each term, the folks thesaurus contains the broader, narrower, related, and equivalent terms, as well as a scope note. Folks thesaurus terms can be used both to broaden and simplify searches. Related Terms, which can be more helpful than synonyms, can make searches more thorough. The use term, or preferred term, reduces the need to search synonyms (Biblioline, 2018).

The folks thesaurus can be used as a guide to other topics. It is helpful whenever users need to construct precise, effective key term searches. It allows terms belonging to the same class to be grouped into hierarchies and makes it possible to establish semantic relationships between these terms and terms from another class. The hierarchical semantic relationship allows the user to have access to narrower or broader concepts within the same class.

Folks thesaurus can be a powerful searching tool. In a regular keyword search, the information system (a general search engine such as Google) finds every document where a word occurs – but not documents where the author used a different word with the same meaning. For example, a search for the word "employees" will find documents with the word "employees" but not documents with the words "employee", "laborer", "laborers",

"stakeholders", "worker", "workers", etc. The folks thesaurus is a way around this problem. By using the folks thesaurus the user can locate the correct term/descriptor for his/her topic and use that to look up information resources. If, for example, the folks thesaurus of a search engine indicates that the term for "employees" is "personnel," doing a thesaurus search using "personnel" will find all documents about that group – even if the author uses the word "employees." It will also get rid of documents which mention employees, laborers, workers, etc. but are not really about them. Essentially the folks thesaurus is a focusing tool. It helps to find all the documents on a topic while at the same time excluding the irrelevant documents.

Applications and implications of Folks Thesauri

Despite the complexity of its development, a folks thesaurus offers many advantages in comparison with an instant's autocomplete suggestions (e.g., “Google Suggest” or “Autocomplete” on the Google search engine). For instance, the efficiency of the selection of terms is improved and recurrent data are eliminated by the hierarchical and associative structure.

By standardizing information that is entered onto a search engine database, it is easier to search documents and retrieve the required output. The use of a folks thesaurus allows the retrieval of information created by someone else, it also allows users to access and retrieve semantic data in the Semantic Web environment. A folks thesaurus can be used as a tool for the Semantic Web.

The simplest way to ensure that the information is consistent, is to use a wordlist of suggested terms. This is simply an alphabetical list of accepted terms used to control the information recorded within a search engine database. However, a wordlist does not allow the user to create relationships between the terms. But a folks thesaurus allows the user to create and browse semantic relationships between the preferred terms. Using a thesaurus structure, can greatly aid information retrieval, applying a recommender system and autocomplete. It also reduces the number of searches required to retrieve information from a search engine, saving the time of the user (Noruzi, 2008). A folks thesaurus remind the user of semantically related ideas that might be valuable in searching.

The information retrieval area has changed dramatically in recent years, followed by the appearance of recommender systems and Semantic Web technologies, with the immense increase in availability of searchable full-text and the increasing availability of powerful search engines (e.g., Google and Google Scholar) for searching the scholarly documents. It is reasonable to ask whether there is any place left for folks thesauri in the Semantic Web information retrieval system. It seems that there is a place for folks thesauri.

Milstead (1998) stated that traditional thesauri must change in order to continue to be of value in the new information retrieval environment.

A true folks thesaurus has equivalence relationships, but it also supports other kinds of relationships, such as genus-species, and provides navigation assistance by means of scope notes and other aids. In other words, a folks thesaurus is a tool designed to aid users in finding their way around a vocabulary database of a search engine. In addition to its use as an authority for the terms used in searching and browsing the search engine database, it recommends terms that the user might not even have considered. Therefore, folks thesauri can facilitate access to related documents on the Web.

Use of Folks Thesauri in Web Information Retrieval

The primary purpose of a folks thesaurus is to aid in web information retrieval, which may be achieved in various ways. This primary use may be achieved by using the folks thesaurus in the automatic indexing of a search engine database, and later in its searching. Secondary purposes include aiding in the general understanding of a subject area, providing 'semantic maps' by showing inter-relations of concepts, and helping to provide definitions of terms, and the support of computer-assisted indexing (Nestel et al., 1992; Aitchison, Gilchrist, & Bawden, 2000, p. 1). Based on information retrieval utility, there are four ways that the folks thesaurus may be used:

1. Folks thesaurus used both in indexing and in searching;
2. Folks thesaurus used in indexing, but not searching;
3. Folks thesaurus used in searching, but not indexing (like the "Google Suggest" or "Autocomplete"); and
4. Folks thesaurus used in neither.

The folks thesaurus is used for searching, but not for indexing: the 'searching thesaurus'. The role of the folks thesaurus here is usually to assist in the searching of a free-text database (e.g., Google) by suggesting additional search terms, especially synonyms and narrower terms. The 'suggestion' may be done explicitly, by offering the terms to the user for their choice, or automatically: this process is generally referred to as 'query expansion' (Aitchison, Gilchrist, & Bawden, 2000, p. 2-3).

Thesauri actually have a place at both ends of the information access process, at both storage and retrieval, for example in ERIC, Embase, Ovid, and ProQuest. The amount of electronically accessible full-text is so immense, and is growing so fast, that web users need all the help they can get in accessing it.

Searching thesauri are often somewhat different in nature from the 'traditional' thesauri, especially in providing a much wider set of terms as an 'entry vocabulary'; conversely, it has been demonstrated that the richness of semantic structures of different kinds in the

traditional thesaurus may be of particular value (Jones et al., 1995). Techniques of construction may also differ, with greater use of automatic and semi-automatic construction techniques, and techniques based on discourse analysis from a cognitive viewpoint (Aitchison, Gilchrist, & Bawden, 2000, p. 3).

In a folks thesaurus, users can assign terms or keywords to other terms or keywords. In this way, hierarchical relationships with superterms and subterms are defined. From these hierarchies, one can derive tree structures like those of known thesauri. Most of the terms can be connected to a selected main term that is superordinated to all other terms (i.e., top term). In a folks thesaurus, terms are connected more flexible with less strict semantics.

While current keyword search engines (e.g., Google, Yahoo, and Bing) may help locate relevant data, they do not take advantage of the semantic relationships between concepts and terms. In this way, folks thesaurus is a collection of terms along with some structure or semantic relationships between them, which can be used at retrieval. A folks thesaurus may work behind the scenes much of the time; while users should certainly have access to any available vocabulary aids if they want them, general search engines (e.g., Google, Yahoo, and Bing) need to redesign their user interfaces so that users interact directly with the folks thesaurus to any greater extent than they wish or need to.

A folks thesaurus can become the basis of a more extensive semantic network or semantic social network, providing information not just on what terms are used by the folks, but on how they are used within the system (e.g., Google). A folks thesaurus is a networked collection of the folks terms. It can use wiki-based softwares, effectively using hyperlinks and hierarchical structure.

As for the role of the folks thesaurus in searching full-text systems, Milstead (1997) believes that in the future, thesauri will be used more in retrieval than in input. Thesauri are used in online databases as both indexing and search tools (Blocks, Binding, Cunliffe, & Tudhope, 2002). Thesauri have tended to be underused in database searches, due to their being frequently unavailable to users. This may change when the work of the folks thesaurus takes place increasingly behind the scenes, or when the user is able to interact with the thesaurus more easily. Where a thesaurus has not been used for indexing (i.e., search engine indexing), a folks thesaurus designed specifically for searching may be available and useful (like the Google Suggest). The exact nature of the search folks thesaurus and how its features differ from those of the traditional thesaurus remain to be fully clarified. It seems that other types of ontology-based relationships are needed, for example, the *generic relation* (is-a), the *subsumption relation* (is-a-superclass-of, the converse of is-a, is-a-subtype-of or is-a-subclass-of), etc.

Web search engines can produce better results by taking advantage of the presence of folks thesaurus or controlled vocabulary of search terms, including related words. In a Boolean system the chances of retrieving relevant documents that do not happen to contain the words of the search query are improved, though precision is not helped unless the search is specifically limited to controlled vocabulary terms (Milstead, 1998).

Some search engines and other knowledge management software suites have built-in corporate taxonomies or 'knowledge structures' that may be automatically generated or manually 'customized'. Starr (1999) describes a number of such tools that have features similar to a search thesaurus, such as clusters of equivalent terms, inter-term relationships, and hierarchies bearing resemblance to tree structures.

Search engines may help the user by including, narrower terms automatically (or at the request of the user). If folks thesauri, and in particular, the semantic relationships within folks thesauri are to be used for web information retrieval, a consistent and well-understood interpretation or semantics is required for the relationships. This is particularly important if query expansion is to be automated (Bechhofer & Goble, 2001).

Over the years there have been proposals for folks thesauri -end-user thesauri- designed specifically to facilitate searching. A folks thesaurus -end-user oriented thesaurus- differs from a traditional thesaurus in two primary ways: its term inclusion and organization, and its displays. It is designed to reflect and organize the total specialized vocabulary of users, rather than to provide a limited list of authorized terms. It gives more information about the scope of terms, and its displays are designed around the way in which users approach information (Bates, 1990; Anderson & Rowley, 1992; Milstead, 1998; Shiri, Revie & Chowdhury, 2002; Sanatjoo, 2007).

Thesauri, originally designed to facilitate consistent analysis of documents at input to an information retrieval system, are already well on their way to becoming vital retrieval tools as well (Milstead, 1998). In fact, folks thesauri may be used more at retrieval than at input. A folks thesaurus can be understood as a kind of Knowledge Organization System (i.e., a search thesaurus) (Hjørland, 2016).

Conclusions

This study provides guidelines for constructing folks thesaurus: formulating the descriptors, establishing relationships among terms, and effectively presenting the information. A good folks thesaurus program can automatically post the reciprocals of relationships and check for consistency. A folks thesaurus can be designed based on the wiki software. It evolves from bottom to top. Folks thesauri take high advantage of the

online electronic medium. Good folks thesaurus management ensures that the thesaurus remains relevant and usable over time. A folks thesaurus grows and evolves over time.

The folks thesaurus can play a key role in enhancing searches in the Semantic Web search engines. To work effectively, the folks thesaurus needs to be designed in accordance with good thesaurus principles and with the user in mind (user warrant). A collaboratively developed folks thesaurus can be used for indexing and searching the Web. It is a new method of information retrieval that combines thesauri and collaborative social web environments. More research is needed to determine what modifications are most likely to make folks thesauri more useful in full-text systems, whether in indexing or searching.

References

Aitchison, J., Gilchrist, A., & Bawden, D. (2000). Thesaurus construction and use: a practical manual. Fourth edition. London: Aslib.

Anderson, J.D., & Rowley, F.A. (1992). Building end-user thesauri from full-text. In Barbara H. Kwasnik and Raya Fidel, eds. *Advances in Classification Research*, Volume 2; *Proceedings of the 2nd ASIS SIG/CR Classification Research Workshop*, October 27, 1991. Medford, NJ: Learned Information, 1992. p. 1-13.

Bates, M.J. (1990). Design for a subject search interface and online thesaurus for a very large records management database. In: D. Henderson (ed.), *Proceedings of the 53rd American Society for Information Science Annual Meeting*, Medford, NJ: Learned Information, 1990. pp. 20-28.

Bechhofer, S., & Goble, C. (2001). Thesaurus construction through knowledge representation. *Data & Knowledge Engineering*, 37(1), 25-45.

Biblioline (2018). Thesaurus. Retrieved December 12, 2018, from http://nes.biblioline.com/help/general/eng/THESAUR.htm

Blocks, D., Binding, C., Cunliffe, D., & Tudhope, D. (2002). Qualitative evaluation of thesaurus-based retrieval. *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, (ECDL 2002), September 16-18, 2002, Heidelberg, Springer-Verlag Berlin, pp. 346-361.

Hjørland, B. (2016). Does the traditional thesaurus have a place in modern information retrieval? *Knowledge Organization*, 43(3), 145-159.

Jones, S. et al. (1995). Interactive thesaurus navigation. Intelligence rules OK? *Journal of the American Society for Information Science*, 46 (1), 52–59.

Milstead, J.L. (1997). Thesaurus in a full-text world. *In:* Cochrane, Pauline Atherton and Eric H. Jones (eds.), *Visualizing subject access for 21st century information resources*. *Proceedings of the 1997 Clinic on Library Applications of Data Processing, 2–4 March 1997.* Urbana-Champaign, Illinois: Illinois University at Urbana-Champaign, Graduate School of Library and Information Science, 1998, pp. 28–38.

Milstead, J.L. (1998). Use of thesauri in the full-text environment. Based on a paper presented at the *34th Clinic on Library Applications of Data Processing* (Cochrane & Johnson, 1998).

Noruzi, A. (2004). Application of Ranganathan's laws to the Web. *Webology*, 1(2), Article 8. Retrieved December 12, 2018, from http://www.webology.org/2004/v1n2/a8.html

Noruzi, A. (2007). Folksonomies: Why do we need controlled vocabulary? *Webology*, 4(2), Editorial 12. Retrieved December 12, 2018, from http://www.webology.ir/2007/v4n2/editorial12.html

Sanatjoo, A. (2007). *Improvement of thesaurus design - through a work-task oriented methodology*. Doctoral dissertation, Department of Information Studies, Royal School of Library and Information Science, Denmark. Retrieved December 12, 2018, from https://curis.ku.dk/portal/files/163188274/985.pdf

Shiri, A. A., Revie, C., & Chowdhury, G. (2002). Thesaurus-enhanced search interfaces. *Journal of Information Science*, 28(2), 111–122.

Starr, J. (1999). Content classification: leveraging new tools and librarians' expertise. *Searcher*, 7 (9), 10-24.

Voss, J. (2006). Collaborative thesaurus tagging the Wikipedia way. *Wikimetrics research papers*, 1(1). Retrieved December 12, 2018, from http://arxiv.org/abs/cs.IR/0604036

Bibliographic information of this paper for citing:

Noruzi, Alireza (2018). "Folks thesauri or search thesauri: Why semantic search engines need folks thesauri?" *Webology*, 15(2), editorial 26. Available at: http://www.webology.org/2018/v15n2/editorial26.pdf