

Evaluating Algorithms for Creation of Hierarchical Category Systems

Megan Richardson
Mathematical Sciences/Linguistics
New Mexico State University
merichar@nmsu.edu

Graduate Mentor: Preslav Nakov
Faculty Advisor: Dr. Marti Hearst

July 6, 2006

Summer Undergraduate Program in Engineering Research
at Berkeley – Information Technology (SUPERB-IT) 2006



Department of Electrical Engineering and Computer Sciences
College of Engineering
University of California, Berkeley

Evaluating Algorithms for Creation of Hierarchical Category Systems

Megan Richardson

Abstract

Hierarchical faceted category systems for search and browsing have been shown to be useful for large information collections. Castanet is an algorithm for automatically generating hierarchical faceted metadata using text descriptions of items in the collection and WordNet, an existing lexical database. The present study evaluates Castanet and two other algorithms through a heuristic evaluation of the category systems they produce, conducted by means of a questionnaire distributed among information architecture professionals and subject domain experts. Participants were asked to rate the systems' meaningfulness, consistency and completeness and to suggest changes to the structures presented. Responses for each algorithm were compared to a baseline, a list of terms occurring frequently in the text descriptions of items in the collections, and to the other algorithms. Castanet rated more highly than the other algorithms, LDA and subsumption, in preference reports. Data were also analyzed to rate the comprehensiveness, coherence and appropriateness of Castanet-generated metadata in order to determine along which parameters Castanet can improve

1 Introduction

1.1 Hierarchical Faceted Search – Flamenco

Faceted hierarchical category structures have been found to be useful in search and browsing (English et al. 02, Yee et al. 03). Facets are orthogonal descriptors used to categorize items in a collection. In Figure 1, “Flavorer, Seasoning,” “Dish,” “Preparation Type” and “Vegetable” are all facets for recipes. Facets typically describe attributes of items, but facets are designed whereas attributes are intrinsic.

The screenshot shows the Flamenco Recipes website interface. At the top, there's a navigation bar with 'Flamenco Recipes' and 'Logged in as m'. Below the navigation bar, there's a search bar and several buttons: 'Save Search', 'History and Settings', 'Return to Search', 'New Search', and 'Log'. The main content area is divided into several sections. On the left, there's a search bar with a 'search' button and a dropdown menu for 'all items' and 'in current results'. Below the search bar, there's a section for 'Refine your search within these categories:' with a list of facets: 'FLAVORER, SEASONING' (group results), 'DISH: all > pasta' (group results), 'PREPARATION TYPE: all > baking', and 'VFGFTARI F'. Each facet has a list of items with counts. For example, under 'FLAVORER, SEASONING', there are 'condiment (13)', 'curry (14)', 'garlic (6)', 'herb (12)', 'sauce (13)', 'spice (3)', 'spread (5)', and 'sweetening (3)'. Under 'DISH: all > pasta', there are 'cannelloni (1)', 'dumplings (7)', 'lasagna (3)', 'macaroni (6)', 'noodle (9)', and 'spaghetti (4)'. Under 'VFGFTARI F', there are 'celery (7)', 'greens (5)', 'legume (1)', 'onion (10)', 'pepper (18)', and 'potato (1)'. On the right, there's a section for 'These terms define your current search. Click the x to remove a term.' with three terms: 'DISH: pasta', 'MEAT AND FISH: poultry > chicken', and 'PREPARATION TYPE: baking'. Below this, there's a section for '24 items, grouped by VEGETABLE' with a link to 'view ungrouped items'. Under the 'celery (7)' category, there are several recipe links: 'Chicken & Dumplings - Bulletin Board Recipes - Southern U.S. Cuisine', 'Chicken Fricassee Recipe - Chicken Fricassee with Dumplings', 'Chicken Noodle Casserole - Recipe for Chicken Noodle Casserole', 'Turkey Macaroni Casserole - Recipe for Turkey Casserole Delight', 'Chicken with Drop Dumplings Recipe - Recipe for Chicken with Fluffy Drop Dumplings', and 'Chicken Stew with Cornmeal Dumplings - Recipe for a Chicken Stew Recipe - Recipes'. At the bottom, there's a section for 'greens (5)' with two recipe links: 'Easy Chicken Parmesan - Chicken Recipes - Southern U.S. Cuisine' and 'Chicken Casserole - Recipe for Chicken Casserole with Macaroni'.

Figure 1: Browsing for recipes in Flamenco, using “Dish,” “Meat and Fish,” and “Preparation Type” facets.

Each facet is associated with a set of labels. If the facet is *hierarchical* rather than *flat*, a label may have other labels beneath it in the structure. Selecting a label is equivalent to performing a disjunction over the labels beneath it. Selecting multiple labels means building a conjunction of disjunctions, for example Pasta AND Baking AND Vegetable={Celery OR Greens OR Legume OR...}.

Facets by themselves do not necessarily capture themes, nor do they show relations. Therefore it is not obvious which facets will be helpful in search. Also, it is an open question how best to assign faceted hierarchical descriptors to items automatically (English et al. 02).

Flamenco, the software which produced the interface shown in Figure 1, is a system for creating a search and browsing interface from faceted hierarchical metadata. Flamenco consists of

- a table keyed by facet, listing all facets;
- a table keyed by item, containing all single-valued metadata for each item; and
- for hierarchical facets,
 - a table keyed by facet value, giving the hierarchy level and the path from the root for each value, and
 - a table associating items with facet values (English et. al unpublished).

Flamenco interfaces allows selection of multiple labels, switching between facets, searching over keywords in addition to using facets, and moving up and down in hierarchies. Flamenco produces usable interfaces but requires faceted hierarchical metadata for each item in the collection.

1.2 Metadata Creation – Castanet

The creation of a faceted hierarchical search and browsing system requires that metadata be assigned to items in the collection, specifying their facets. Precisely because these collections are large, hand-coding of this metadata is costly; hence the need for algorithms to assign metadata.

Castanet is an algorithm that creates facets and hierarchies and assigns metadata accordingly. First, frequent terms are taken from the items or their descriptions. Second, hypernym paths are taken from WordNet, a handmade lexical hierarchy, for one sense of each word (<http://wordnet.princeton.edu>). Usually, the sense most common in WordNet is used. Trees are made by combining these paths, amounting to a subset of WordNet corresponding to the subject matter of the collection. These trees are then compressed by

1. eliminating very general categories;
2. eliminating nodes with too few children (this parameter can vary); and
3. eliminating child nodes whose names appear in their parent nodes.

The hierarchical faceted metadata are then assigned according to the collection-frequent terms appearing in each item or its description.

Castanet appears to produce category structures useful for creation of search and browsing interfaces (Stoica & Hearst 04). However, the quality of Castanet's output is difficult to quantify.

2 Related Work – Category System Evaluation

There is no widely accepted means of evaluating category systems for search and browsing. Several metrics have been applied, but each has drawbacks.

Since category system creation has some historical roots in the field of information retrieval (IR), traditional IR evaluation metrics have been applied. These include precision, recall and combinations thereof such as F-measure. (Precision is the proportion of relevant items retrieved to the total number of

items retrieved. Recall is the proportion of relevant items retrieved to the total number of relevant items in the collection. F-measure is the harmonic mean of precision and recall.) Such measurements are highly practical in that they can easily be retaken with each iteration of development of the system; unfortunately, they have not been definitively correlated with usability of the system. Category systems are meant to be used by people for tasks; therefore while these indirect metrics are defensible as cognitive models their validity as such must be established. Furthermore, some of the metrics in this class apply better to clusters than hierarchies, having been developed for use with document retrieval and similar applications where output can meaningfully be regarded as bags of words.

Iterative task-oriented usability testing can directly assess the usefulness of a system. Usability testing involves prototyping a system, finding problems in the usability of the system by observing representative users performing representative tasks with the system, and modifying the system based on these observations. This method, however, is not a science but a tool for development and it can be costly and time-consuming.

IR metrics are intrinsic metrics taken using automatic methods. Usability criteria are extrinsic metrics taken using empirical methods. Table 1 classifies some category system evaluation metrics in terms of whether the methods are *automatic* (*ad hoc*, implemented in software) or *empirical* (*post hoc*, involving surveying or testing with representative human participants), and *intrinsic* (expressing purportedly desirable properties of the system, in practice by comparison to another, good system) or *extrinsic* (assessing the usefulness of system in completion of tasks).

	Automatic Methods	Empirical Methods
Intrinsic Metrics	Comparison to Ideal System <ul style="list-style-type: none"> • precision and recall • average uninterpolated precision [Nanas, Uren & de Roeck 03] • similarity as ratio of pairs common to two hierarchies to total pairs in one hierarchy [Lawrie & Croft 00] • similarity as category distance [Sun & Lim 01] • path length [Lawrie & Croft 03] 	User Evaluation (<i>includes the present study</i>) <ul style="list-style-type: none"> • percent correct relationship between pairs [Sanderson & Croft 99] • understandability [Pirolli 96] • accuracy [Krowne & Halbert 05; Li, Zhu & Ogihara 03] • preferences
Extrinsic Metrics		Usability Testing <ul style="list-style-type: none"> • task completion rate • time to completion • preferences [Sanderson & Croft 99; Wu, Shankar, & Chen 03; English et al. 02; Kumamuru et al. 04]

Table 1: Category system evaluation metrics.

The present study heuristically evaluates three category system algorithms through a survey distributed among information architecture professionals and subject domain experts. The survey presented category systems produced by three algorithms: Castanet, latent Dirichlet allocation (LDA) (Blei et al. 04), and subsumption (Sanderson and Croft 99). Each evaluator assessed Castanet, one of the other algorithms, and a baseline (a list of terms occurring frequently in the collection). This method was considered an acceptable compromise between the ease and rigor of automatic methods and the validity of usability testing.

3 Study Design

The participants were asked to rate aspects of the systems' meaningfulness, consistency and completeness. Participants were also asked about changes they would make to the systems to make them more useful. Data were analyzed to rate the comprehensives, coherence and appropriateness of Castanet-generated category systems in order to determine along which parameters Castanet can improve. The study was also designed to test the hypothesis that Castanet generates category systems that information architects and users prefer to those produced by LDA or subsumption and to a baseline list of frequent terms. The study allowed for both within-subjects analysis (each participant evaluated two systems) and between-subjects analysis.

The questionnaire was first tested with two users, an expert in information architecture and an expert in biomedical informatics. Critiques were solicited of the instructions, cover letter, questions and answer choices, especially concerning their clarity and, for questions, their number.

Two datasets were used: a list of 3,275 journal titles taken from MEDLINE citations, and a collection of 13,000 recipes found on web sites. These collections were previously used in Castanet pilots and found to have fewer ambiguities than other collections. It is important to use participants with interest in the subject matter for studies of search (Borland & Ingwersen 97), so information architects with an interest in cooking were presented with the recipe browsing systems and biologists, doctors, medical students and medical librarians were presented with the biomedical journal browsing system. Participation was solicited on email discussion lists dedicated to the relevant subject matter. Participants were entered in a drawing for a \$100 gift certificate for Amazon.com.

All participants answered questions about their information architecture experience (for the recipes interfaces) or their experience with PubMed or similar systems (for the biomedical interfaces). (PubMed is an interface for searching and browsing MEDLINE information on biomedical journal articles.)

The interfaces were given neutral names: "Pine" for Castanet, "Oak" for LDA, and "Birch" for subsumption. Each participant was asked to evaluate two systems, Castanet and either LDA or subsumption. The presentation order was varied.

Participants were asked about the top-level categories for the algorithm, specifically, whether they would add, remove, rename, merge or split any categories. (See Appendix 1 for the questions asked and the response scales.) Participants were then asked to look at two second-level categories and asked whether they would add, promote to top-level, move or remove any subcategories, and whether the category met their expectations. About each system, participants were asked whether they agree with the following statements:

- "Overall, these categories are meaningful."
- "Overall, these categories describe the collection's contents in a systematic way."
- "These categories capture the important concepts for this collection."

After seeing two systems, participants were asked if they would like to use each of the systems to assist them in creating a web site for browsing the collection, and whether they would like to use a simple list of frequent terms from the collection.

Throughout, subjective ratings were reported using a four-point scale and specific information was recorded on changes the participants would make to improve the category system. Modes of scaled responses were found. Preference information for the three algorithms and the baseline were compared. Responses concerning overall ratings of the systems were compared across the algorithms and chi-

squared tests were conducted to establish the significance of differences found. Scales were treated as categorical responses. Inter-rater agreement was calculated using Kendall's coefficient of concordance.

The completeness, coherence, and appropriateness of categories created by Castanet was assessed by looking at

- the extent to which participants would add categories and
- the extent to which participants found the system captured the important concepts for the collection (for completeness);
- the extent to which participants would rename, merge, split or move categories and
- the extent to which the categories were considered systematic and meaningful (for coherence);
- the extent to which participants would remove categories and
- the extent to which subcategories met participants' expectations based on the labels (for appropriateness).

Specific suggestions relating to the above and general comments offered by evaluators were saved for consideration in improving Castanet.

4 Results

Responses to the questionnaire were accepted for one week. There were 49 evaluations of Castanet, 28 evaluations of LDA, and 17 of subsumption. These figures differ from figures (below) for comparative evaluations because some participants dropped out after evaluating individual systems but before finishing the comparison on the last page of the questionnaire. For all parts of the evaluation, in all conditions, levels of participation were sufficient for heuristic evaluation (Nielsen & Landauer 93).

4.1 Comparative Evaluation

Overall, Castanet was preferred to the other algorithms and the baseline list of common terms, with 76% of Castanet raters saying they might or would definitely use the Castanet category system, as contrasted with four percent for LDA, 31% for subsumption, and 63% for baseline (see Figure 2). Differences in results for different presentations (which other system was shown with Castanet, in which order) were within one percent, with one exception: six percent more users would "definitely not" use Castanet for biomedical journals when shown with subsumption rather than LDA.

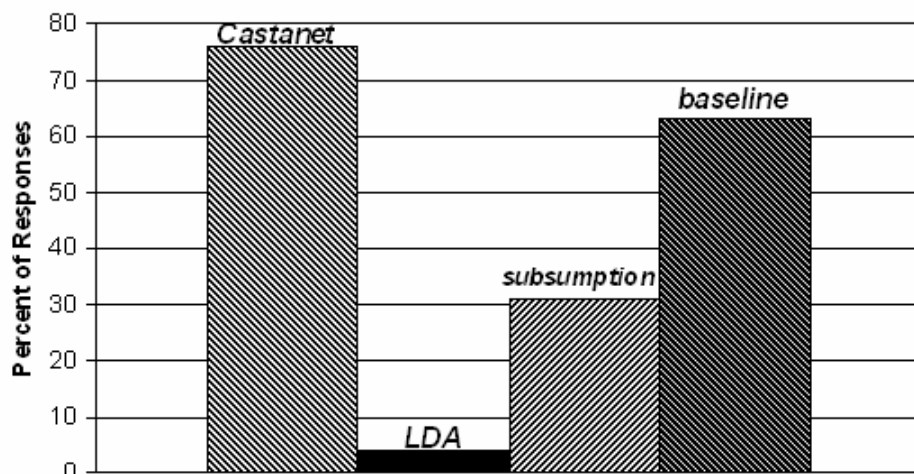


Figure 2: Percent of participants who "might want to use" or "would definitely use" each system.

The mode response for both Castanet and baseline was “I might want to use this system in some cases,” however, more respondents “would definitely use” Castanet. Preference information is given in Table 2. Differences were significant ($X^2 = 169.0486$, $df = 9$, $p \leq 0.001$); however, results concerning subsumption in particular may not be meaningful due to the lower number of complete responses.

	No, definitely not	Probably not	Yes, I might want to use this system in some cases	Yes, I would definitely use this system
Would you like to use [Castanet]? (N=40)	10%	15%	58%	18%
Would you like to use [LDA]? (N=24)	75%	21%	4%	0%
Would you like to use [subsumption]? (N=16)	31%	38%	25%	6%
Would you like to use a list of frequent terms [baseline]? (N=40)	15%	23%	55%	8%

Table 2: Preference responses. Percentage of participants giving the mode response in bold.

4.2 Overall Direct Evaluation

Certain prompts sought general impressions of a system without comparison to other systems. Those prompts asked for evaluators’ degree of agreement with the following statements:

- “Overall, these categories are meaningful,”
- “Overall, these categories describe the collection’s contents in a systematic way,” and
- “These categories capture the important concepts for this collection.”

In every condition, the mode response to these prompts was “agree somewhat” for Castanet. For LDA and subsumption, the mode response for all three questions was “strongly disagree,” except for subsumption in the biomedical journals conditions where the mode response was also “agree somewhat.” Mode responses are given in Table 3. Less common responses are given in Appendix 2.

	Castanet		Subsumption		LDA	
	Biomedical N=16	Recipes N=30	Biomedical N=8	Recipes N=9	Biomedical N=8	Recipes N=9
These categories are meaningful	Agree somewhat (81%)	Agree somewhat (37%)	Agree somewhat (50%)	Strongly disagree (89%)	Strongly disagree (75%)	Strongly disagree (79%)
These categories describe the collections contents in a systematic way	Agree somewhat (81%)	Agree somewhat (37%)	Agree somewhat (50%)	Strongly disagree (78%)	Strongly disagree (62%)	Strongly disagree (63%)
These categories capture the important concepts	Agree somewhat (81%)	Agree somewhat (57%)	Agree somewhat (38%)	Strongly disagree (67%)	Strongly disagree (75%)	Strongly disagree (79%)

Table 3: Mode responses to direct evaluation prompts. Percentage of participants giving the mode response in parentheses.

4.3 Comprehensiveness, Coherence and Appropriateness of Castanet Output

In general, Castanet received the top rating where specific questions were asked about improvements to particular categories, indicating that participants would not make changes. The system did especially well in measures of coherence and appropriate depth. Yet where general impressions were sought Castanet fared less well. For example, when asked “Are there any categories you would split into two, three, or more categories?” participants typically responded “No, none,” but when asked “Did this category match your expectations?” participants typically responded “Somewhat.” Comprehensiveness, coherence and appropriateness response information is given in Appendix 2 and summarized in Table 3. Inter-rater agreement was significant at $W = 0.4309$ ($X^2 = 63.7695$, $df = 4$, $p < 0.0001$). See Appendix 1 for rating scales for each question.

	Best Rating e.g., "Strongly agree"	Intermediate Ratings e.g., "Agree somewhat" e.g., "Disagree somewhat"	Worst Rating e.g., "Strongly disagree"	
Overall Impressions:				
Systematicity	11%	62%	15%	11%
Meaningfulness	15%	59%	20%	7%
Intuitivity	23%	26%	38%	14%
Important Concepts	6%	69%	18%	7%
Specific Changes:				
Completeness	46%	32%	15%	8%
Correct Naming	45%	29%	22%	5%
Differentiation	49%	29%	20%	2%
Category Coherence	81%	14%	3%	2%
Correct Depth	74%	22%	4%	0%
Correct Placement	63%	26%	9%	3%
Scope	54%	31%	9%	6%

Table 3: Summary of non-comparative ratings for Castanet. Percentage of participants giving the mode response in bold.

Where Castanet did receive suboptimal scores on questions concerning specific changes, the questions pertained to removing, merging and renaming categories. (Suboptimal scores are in bold and italics in Appendix 2.) Evaluators found top-level categories (in recipes condition) and subcategories (in biomedical journals condition) that they would remove. Evaluators would also merge some top-level categories in recipe condition. Participant's comments relating to removing categories frequently mention a surplus of categories, categories that could be merged and categories that are ambiguous, too general or too broad.

5 Conclusions and Future Work

In general, Castanet performed slightly better than baseline and considerably better than other algorithms. Inter-rater reliability suggests that the chosen evaluation method is meaningful. Since production of the metadata used in the study was fully automatic, the results obtained for Castanet suggest that fully automated metadata competitive with handmade metadata is within reach. However, because results concerning subcategories suggest that they are very well-formed, Castanet's shortcomings are difficult to diagnose.

High ratings for Castanet are less pronounced for the biomedical journals than the recipes (see Appendix 2). This may be because making judgments about biomedical journals is more difficult. It may also be that Castanet performs less well in the biomedical domain. Castanet's WordNet-derived hierarchies may be insufficient because the narrow definitions of medical terminology limit the possible paths through WordNet. Selection of paths using senses rather than words might therefore improve Castanet. Scores concerning merging categories, and suggestions to merge found in comments on category removal, also suggest that a mechanism for handling synonymy would improve performance. Alternatively, the use of a simplified version of WordNet featuring combined synsets could be tested (Mihalcea & Moldovan 01, Stoica & Hearst 04). Examination of the suboptimal scores on questions concerning specific changes suggests that Castanet may be creating too many categories, so taking synonymy into account could improve matters. One must guard against creation of overly broad categories, however.

Sense-disambiguation functionality might also help if, as appears to be the case from participant comments, suggestions to remove categories reflect the presence of some overly broad categories. This should also improve performance if suggestions to remove categories reflect the presence of misfit concepts in the categories. We might also experiment with the number of children a node must have in order not to be pruned.

Evaluations using displays other than Flamenco might be conducted to eliminate the possibility that user preferences found in the present study were preferences for the combination of Castanet and Flamenco, rather than for Castanet generally. Evaluations with finer response scales, while potentially less meaningful if few responses are received, could otherwise better illuminate the extent of Castanet's imperfections, which may be slight. Anchored scales would allow for use of different statistical analyses which might be illuminating. Also, more specific responses should be required (rather than optional) in future studies since high specificity in how a system misses the mark might resolve the difficulty of suboptimal systems rating highly across diagnostically informative measures.

Research into automatic category creation in general stands to benefit from empirical testing to correlate automatic evaluation methods with usability, and from the creation of gold standard datasets for use with those methods.

References

- D. Blei, T. Griffiths, M. Jordan and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, 16. 2004.
- P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. In *Journal of Documentation*, 53(3). 1997.
- J. English, M. Hearst, R. Sinha, K. Swearingen and K. Yee. Flexible search and navigation using faceted metadata. Unpublished.
- J. English, M. Hearst, R. Sinha, K. Swearingen and K. Yee. Hierarchical faceted metadata in site search interfaces. In *Communications of the ACM*, 45(9). September 2002.
- A. Krowne and M. Halbert. An initial evaluation of automated organization for digital library browsing. In *Proceedings of JCDL '05*. 2005.
- K. Kumnamuru, R. Lotlikar, S. Roy, K. Singal and R. Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *WWW2004*, May 2004.
- D. Lawrie and W.B. Croft. Discovering and comparing topic hierarchies. In *Proceedings of RIAO*. 2000.
- D. Lawrie and W.B. Croft. Generating hierarchical summaries for web searches. In *Proceedings of SIGIR '03*. 2003.
- T. Li, S. Zhu and M. Ogihara. Topic hierarchy generation via linear discriminant projection. In *Proceedings of SIGIR '03*. 2003.
- R. Mihalcea and D. Moldovan. eXtended WordNet: Progress Report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*. June, 2001.
- N. Nanas, V. Uren and A. de Roeck. Building and applying a concept hierarchy representation of a user profile. In *Proceedings of SIGIR '03*. 2003.

J. Nielsen and T.K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of ACM/IFIP INTERCHI '93*. 1993.

P. Pirolli, P. Schank, M. Hearst and C. Diehl. Scatter/Gather browsing communicates the topic structure of a very large text collection. In *Proceedings of SIGCHI '96*. 1996.

M. Sanderson and B. Croft. Deriving concept hierarchies form text. In *Proceedings of SIGIR '99*. 1999.

E. Stoica and M. Hearst. Nearly-automated hierarchy creation. In *Proceedings of HLT-NAACL '04*. 2004.

A. Sun and E.-P. Lim. Hierarchical text classification and evaluation. In *Proceedings of the 1st IEEE International Conference on Data Mining*. 2001.

Y.B. Wu, L. Shankar and X. Chen. Finding more useful information faster from web search results. In *Proceedings of CIKM '03*. 2003.

K. Yee, K. Swearingen, K. Li and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of CHI '03*. 2003.

Appendix 1: Questionnaire – Pine (Castanet), with comparison to Oak (LDA)

Category Structures Survey (Pine)

Please take your time looking at the categories. We are interested in how complete and coherent these category structures are. We will be asking you about categories you think should be added to, removed from, or changed in each of two category schemes.

1. Would you add any top-level categories?
(No, none Yes, one or two Yes, a few Yes, many)
 2. If "yes," please list some top-level categories you would add. (Optional)
 3. Would you remove any top-level categories?
(No, none Yes, one or two Yes, a few Yes, many)
 4. If "yes," please list some top-level categories you would remove. (Optional)
 5. Are there any categories you would keep, but rename?
(No, none Yes, one or two Yes, a few Yes, many)
 6. If "yes," please list some top-level categories you would rename, along with the new names you would give them. (Optional)
 7. Are there any categories you would merge together?
(No, none Yes, one or two Yes, a few Yes, many)
 8. If "yes," please list some groups of top-level categories you would combine. (Optional)
 9. Are there any categories you would split into two, three, or more categories?
(No, none Yes, one or two Yes, a few Yes, many)
 10. If "yes," please list some categories you would split. (Optional)
- Think about what you would expect to find under "Bread." Look at the subcategories under "Bread."
11. Did this category match your expectations?
(Yes, very closely Yes, mostly Somewhat No, not at all)
 12. Would you add any subcategories to this category?
(No, I would not add any Yes, one or two Yes, a few Yes, many)
 13. If "yes," please list some subcategories you would add. (Optional)
 14. Would you promote any subcategories to top-level?
(No, I would not promote any Yes, one or two Yes, a few Yes, many)
 15. If "yes," please list some categories you would promote. (Optional)
 16. Would you move any subcategories to a different existing top-level category?
(No, I would not move any Yes, one or two Yes, a few Yes, many)
 17. If "yes," list some subcategories would you move, and the top-level categories to which you would move them. (Optional)
 18. Would you remove any subcategories entirely?
(No, I would not remove any Yes, one or two Yes, a few Yes, many)
 19. If "yes," please list some subcategories you would remove. (Optional)

Now think about what you would expect to find under "Cooking." Now look at the subcategories under "Cooking."
[same questions as above]

Please look at a few more categories and subcategories. Then consider how much you agree or disagree with the following three statements about the Pine category system as a whole.

29. Overall, these categories are meaningful.

(Strongly disagree Disagree somewhat Agree somewhat Strongly agree)

30. Overall, these categories describe the collection's contents in a systematic way.

(Strongly disagree Disagree somewhat Agree somewhat Strongly agree)

31. These categories capture the important concepts for this collection.

(Strongly disagree Disagree somewhat Agree somewhat Strongly agree)

You may now close the browser window in which Pine appears.

Overall, would you like to use these tools to assist you in creating a website or other search and browsing interface? Please answer for each of the two systems you looked at. Also consider whether you would like to use a simple list of frequent terms from the collection.

1. Would you like to use Oak?

(No, definitely not Probably not Yes, I might want to use this system in some cases Yes, I would definitely use this system)

2. Would you like to use Pine?

(No, definitely not Probably not Yes, I might want to use this system in some cases Yes, I would definitely use this system)

3. Would you like to use a simple list of frequent terms from the collection of recipes?

(No, definitely not Probably not Yes, I might want to use this system in some cases Yes, I would definitely use this system)

Appendix 2: Comprehensiveness, Coherence and Appropriateness Measures for Castanet

	Best Rating	Intermediate Ratings	Worst Rating	
COMPLETENESS:				
Top Level:				
“Would you add any top-level categories?” (recipes)	33%	27%	30%	10%
(biomedical)	63%	38%	0%	0%
Subcategory A:				
“Would you add any subcategories to this category?” (recipes)	47%	30%	17%	7%
(biomedical)	38%	31%	13%	19%
Subcategory B:				
“Would you add any subcategories to this category?” (recipes)	40%	40%	17%	3%
(biomedical)	56%	25%	13%	6%
COMPLETENESS AVERAGES	46%	32%	15%	8%
Overall:				
“These categories capture the important concepts for this collection.” (recipes)	7%	57%	23%	13%
(biomedical)	6%	81%	13%	0%
‘IMPORTANT CONCEPTS’ AVERAGES	6%	69%	18%	7%
COHERENCE:				
Top Level:				
“Are there any categories you would keep, but rename?” (recipes)	33%	13%	43%	10%
(biomedical)	56%	44%	0%	0%
CORRECT NAMING AVERAGES	45%	29%	22%	5%
“Are there any categories you would merge together?” (recipes)	23%	40%	33%	3%
(biomedical)	75%	19%	6%	0%
DIFFERENTIATION AVERAGES	49%	29%	20%	2%
“Are there any categories you would split into two, three, or more categories?” (recipes)	80%	10%	7%	3%
(biomedical)	81%	19%	0%	0%
CATEGORY COHERENCE AVERAGES	81%	14%	3%	2%
Subcategory A:				
“Would you promote any subcategories to top-level?” (recipes)	90%	3%	7%	0%
(biomedical)	63%	31%	6%	0%
Subcategory B:				
“Would you promote any subcategories to top-level?” (recipes)	63%	33%	3%	0%
(biomedical)	81%	19%	0%	0%
CORRECT DEPTH AVERAGES	74%	22%	4%	0%
“Would you move any subcategories to a different existing top-level category?” (recipes)	63%	23%	13%	0%
(biomedical)	63%	38%	0%	0%

Subcategory B				
“Would you move any subcategories to a different existing top-level category?” (recipes)	87%	10%	3%	0%
(biomedical)	38%	31%	19%	13%
CORRECT PLACEMENT AVERAGES	63%	26%	9%	3%
Overall:				
“Overall, these categories describe the collection's contents in a systematic way.” (recipes)	17%	43%	23%	17%
(biomedical)	6%	81%	6%	6%
SYSTEMATICITY AVERAGES	11%	62%	15%	11%
“Overall, these categories are meaningful.” (recipes)	23%	37%	27%	13%
(biomedical)	6%	81%	13%	0%
MEANINGFULNESS AVERAGES	15%	59%	20%	7%
APPROPRIATENESS:				
Top Level:				
“Would you remove any top-level categories?” (recipes)	23%	27%	27%	23%
(biomedical)	56%	44%	0%	0%
Subcategory A:				
“Would you remove any subcategories entirely?” (recipes)	57%	30%	13%	0%
(biomedical)	38%	56%	6%	0%
Subcategory B:				
“Would you remove any subcategories entirely?” (recipes)	70%	27%	3%	0%
(biomedical)	81%	0%	6%	13%
SCOPE AVERAGES	54%	31%	9%	6%
Subcategory A:				
“Did this category match your expectations?” (recipes)	40%	30%	27%	3%
(biomedical)	13%	19%	63%	6%
Subcategory B:				
“Did this category match your expectations?” (recipes)	27%	37%	23%	13%
(biomedical)	13%	19%	38%	31%
INTUITIVITY AVERAGES	23%	26%	38%	14%