Chris Krause

LIBR 202 Midterm

1. Aggregation, discrimination, and disambiguation are means of filtering, sorting and differentiating data in order to return relevant records in response to a query. These rules structure and shape our information retrieval systems so that the mass of data which they hold can be accessed in a coherent, deliberate, human way. Without such restrictions and parameters databases and information systems become functionally useless, like a novel without punctuation, page number or grammar; raw data without order or rules to sort it. This question speaks to the issue of findability: is it possible to find the record that I have in mind using the rules available to me? Aggregation, discrimination, and disambiguation are some of these common rules.

Data aggregation is the most basic function of an information system: it involves a basic rule to return a composite of relevant records from a database or series of databases. When we search for "cars" the search rule returns records with the quality of being a car, forming an aggregate result. This aggregate can further be restricted by means of discrimination and disambiguation rules.

Discrimination refers to the capacity for an IR system to differentiate between records based upon differences in attributes. We might search among aggregated records of houses to find a specifically blue house, discriminating against the vast majority of other records which include houses of irrelevant attributes. Discrimination is often handled by simple Boolean operators: House AND blue, NOT apartment OR studio.[1] Keywords, tags and controlled vocabulary lists, when attributed to records, also contribute to efficient, discriminating queries.[2] Discrimination allows a searcher to sort records by class, genus and species. Discrimination is a critical aspect of IR in the sense that without it, we would be unable to locate exactly what we are looking for and instead be bombarded with a mass of irrelevant information. Poor discrimination performance was the downfall of many of the earliest search engines.

Ambiguity refers to confusion between what is requested and what is returned from an IR system. Ambiguous natural language has been recognized as having a negative effect on the performance of such systems, often returning records which are not relevant. This is often referred to as the polysemy problem.[3] Disambiguation rules are applied to avoid this dysfunction and are common in our everyday IR life. When we type something incorrectly into Google, Google asks us if we meant something else and offers a handy little link to bring us to a relevant record. But incorrect spelling word sense is only the most superficial form of disambiguation. A system with sophisticated heuristics can interpret the intent behind a searcher's query and direct him or her to something content relevant.[4] In this sense it is possible to enter in an anachronistic term or cryptic question into Google and the engine will offer an alternative in order to help the user find what the disambiguation rules estimate the searcher is actually looking for.

2. Metadata is "data about data," additional information about the records of a database attributed for purposes of cataloging, sorting and user interface; metadata is not the data itself, but additional data that has been created and assigned to it in order to increase findability.[5] There are three primary metadata forms: descriptive metadata, which we will be most interested in for purposes of this question, administrative metadata and structural metadata.

Descriptive metadata is "information describing the intellectual content of the object,"[6] commonly visible to users who utilize information systems. This sort of metadata is involved with helping a user find a record by use of catalog aids including keywords, summaries and classifications. When we go to a library and find that the collections have been sorted by category and keyword, we are interpreting metadata: a system or professional has gone through the effort of tagging the data so that it is easier to locate. Administrative metadata is involved with providing the information necessary for a

repository to manage the object. [7] Structural metadata is involved with differentiating data into logical and coherent groupings: it is responsible for informing a system that pages of data form a book, have a structure that exists outside of the raw information itself. While descriptive metadata is often visible to the user, the other two types work "behind the scenes" and at the disposal of librarians and other information professionals to ensure that the IR system is functional, efficient and responsive.

The simplest example of descriptive metadata concerns images. What we perceive as a picture of two fighting martial artists has no inherent "fightingness" in it, but rather we as humans attribute that verbose quality to it. To find such an image in a search engine, the image must have descriptive metadata added to it, unless there is some spectacular technology that is able to interpret the verbose content of the image. Google has pioneered in this regard, offering an option in the image search to return "faces only," but is still far from being able to interpret the intellectual nature of the records completely. At its most basic application descriptive metadata is "keywording," attaching descriptors to a piece of data that otherwise lacks the inherent verbose quality of that keyword; it adds a human understanding to the record which increases findability.[8]

3. Full text/natural language, controlled vocabulary and classification are three ways in which subjects can be represented. All three of these methods are aids in the finding and acquisition of data and are alternative rules we can apply to an information system's query, assuming it supports that functionality.

Full text or natural language is a method of querying a database that involves the use of everyday language. We might ask a database a question such as "when was John F Kennedy assassinated" or "how far away is the Earth from the sun?" It is as if the search engine was a live librarian and the searcher was at a reference desk, helping another human being for assistance rather than interfacing with database software. This capacity enables the user to more naturally search for

records and removes limitations imposed by a developer, who might otherwise require the use of specific keywords, Boolean operations or technical language.[9] While this method of subject representation has been utilized since the beginning of IR, it was emerged publically with the advent of Ask Jeeves, a search engine that boasted in marketing materials that it could answer natural questions with relevant records. While the initial software behind Ask Jeeves was anything but spectacular, today all major search engines can accommodate natural language queries, and have advanced heuristics to interpret the intent behind the searcher's queries. More generally, full text subject representation refers to a querying methodology in which a text document is scanned in order to return a specific instance. Google uses this method.

Another way a subject may be represented is through controlled vocabularies. Controlled vocabularies are interpretive lists that sort records based upon content. The most prevalent example is the Library of Congress classification system, which houses records under twenty common themes, ranging from military science to bibliographies. The data is sorted under the appropriate class and may be further delineated by the use of magnet subclasses which lead to more specific content. The downside to this method in contrast to natural language representation is that it is less likely to return records which are specifically relevant. While natural language allows one to locate specific data within records, there is also a risk of irrelevance or poor quality control due to lack of organization under a greater classification scheme.[10] An example query using a controlled vocabulary might be "coats," which will return options for winter coats, spring coats, cotton coats, with appropriate subclasses for men, women, children, size, color etc; the user would navigate a tree of classes and dependant subclasses. In this sense a controlled vocabulary interprets semantically what the user is attempting to find and then directs him or her toward categories of records.

Classification is an attempt to catalog a record on the basis of content type. Historically the most common system has been the Dewey Decimal Classification in public and general purpose libraries and the Library of Congress system in academic libraries.[11] This is the system has a practical application in paper libraries and a nearly anachronistic presence in digital collections. It has the advantage of directing a searcher to patently relevant records, but at the significant expense of findability in regard to specific, keyword queries. A searcher will surely find records on history under sections labeled history, but without a controlled vocabulary this system becomes cumbersome and unwieldy, more useful as a general reference system rather than as a focused means of information retrieval.

4. I believe #3 sufficiently defined what a controlled vocabulary is so herein I will focus on the latter two aspects of the question. Controlled vocabularies become very useful when cataloging vast and complex systems of interrelated data. The prime example of this is in shopping online: it would be tedious if not impossible to locate items we are interested in on a site like Amazon.com if not for controlled vocabulary. We are interested in a comforter for a bed, so we query home décor, which leads us to the different rooms of the house, which accordingly leads us to the bedroom, and from there, to comforters, pillows, beds, dressers, desks, and from there to our desired category, and even still to subclasses including color, size, and every other attribute under the sun. In this fashion we query and refine our search until we find exactly what we had in mind.

While a full text system would allow us to find a specific record such as “MARRIKAS 300TC Egyptian Cotton PillowCase” with ease, what if we don’t know specifically what we are looking for and instead are desirous of the serendipity of an effective IR system to steer us to relevant records? Without a controlled vocabulary to filter out records above the current tree level and to propose subcategories for our delving, we find ourselves lost in such a task. In this sense controlled vocabularies are very useful

for navigating a complex world of data and presenting us with a wide breadth of options, but are not particularly useful in efficiently returning specific, focused queries with good effect.

The process of creating and maintaining a controlled vocabulary is one of the main disadvantages of this method of subject representation, as it requires extensive time investment by professional personnel to either catalog records themselves or to design an interactive system in which users can voluntarily catalog the same records. The first step is to breakdown the data into a tree for accessibility sake in a similar fashion to how organisms are classified in science: every aspect of difference must have its own subcategory and layer of detail. The records must then be attributed in a fashion so that they can be cross-referenced by the database rules (software). In this way the librarian can deliberately design the way in which users navigate the record structure for purposes of streamlining discovery and findability. The weakness of such a system is that it must be maintained in the event of new record types emerging, otherwise it loses its functionality. A particularly crafty team might create a system in which this management work is mostly done by the users so that the professional staff oversees rather than completes all of the cataloging. Many mainstream websites have adopted this model, from Wikipedia to EBAY. While the initial effort of developing such a system is high, the management cost is minimal, wherein professions need only interfere on the occasion of outstanding circumstances.

5. Boolean logic is at the heart of digital electronics and computer systems, referring to complete system of rules which may be utilized to sort, exclude and include records into search results. At its most primitive Boolean logic is used in the operation of basic circuits, informing a system to transmit or not to transmit data (1s and 0s), and so forming the foundation of modern computer science. In library science

we use the Boolean operators (rules) to more efficiently locate records and to narrow down a large search aggregate to a specific query.[12]

The most common Boolean operators are:[13]

AND: Include records which must include multiple words. “Oranges AND tomatoes” queries records which feature both oranges and tomatoes.

OR: Include records which may or may not include multiple words. “Oranges OR tomatoes” queries records which feature oranges or tomatoes.

NOT: Includes records which must not include word(s). “Oranges NOT tomatoes” queries records which feature oranges, but must not include mention of tomatoes.

( ): Can be used mathematically with the basic Boolean operators to perform advanced queries. “(Tomatoes OR oranges NOT celery) AND (Low sodium or healthy)” queries records which may include tomatoes or oranges, but must not include celery, and those initially sorted records also must have low sodium or be healthy.

No operator: Includes records which match the exact query and ignore all operators.

WITHIN: Includes records within a certain word radius. “Oranges WITHIN 3 tomatoes” queries records which feature the word oranges within three word spaces of tomatoes.

NEAR: See WITHIN.

BEFORE: Includes records that feature the word coming in a particular order in regard to another word. “Tomatoes BEFORE oranges” queries records which feature the word tomatoes coming in the text before the word oranges.

AFTER: See BEFORE.

: Includes records that contain a common feature in exception of the wildcard. “Astro” queries records which feature the root “astro” in addition to anything else: i.e. astrology, Astroturf, astronomy etc. Can also be used within words to query records for variations on spelling: i.e. “ast*o” queries records which contain words beginning with “ast” and end with “o.”

?: A single character wildcard also used for spelling, restricted to the number of instances in the query. “???fare” queries records which include the suffix “fare” and also include a prefix of exactly three characters: i.e. “welfare” “warfare” “busfare.”

Practical example:

Joseph Blowh is querying a database about Confederate prison conditions during the American Civil War. He begins with “Prisons” and finds millions of records are returned. He might create a new query “Civil War AND prisons,” which returns tens of thousands of results. Mr. Blowh is getting closer, but his query is still too general, and must refine his search in order to find a good citation. The good Mr. Blowh modifies his search to “Civil war AND prisons AND (Confederate OR CSA OR Confederate States of America).” That query was very close, but there are still a few hundred results, and at least half of them are about general prison information during the civil war. Many records include information about the Union prisons as well as the Confederate prisons. He modifies his search again: “Civil war AND prisons AND (Confederate OR CSA OR Confederate States of America) NOT (Union prisons or Federal prisons or Northern prisons).” He now only has a few dozen records to search through, and all of them are relevant.

6.Pre-coordination and post-coordination speaks to the issue of how subject headings are constructed, maintained and utilized. Pre-coordinated headings are the traditional method of organizing subject representation, used by the Library of Congress. [14] This method involves the data being cataloged and sorted into complex topics of classification before the end-user has access to it, either as it applies to a controlled vocabulary or as it is decided by a cataloger. In both instances, complex topic data (ranging from "the effect of violence on the family" to "World War II military strategies") is stored in the record.[15] In this fashion pre-coordinated headings are often more expressive than post-coordinated ones and thus return more relevant records when queried than the Boolean operator dependant latter, although at the cost of manual indexing. In this approach efficient cataloging "consists basically of finding the best match between the work being cataloged and the available pre-combined headings."[16] Post-coordination involves the assigning of single concept headings to bibliographic records, "allowing the combination to take place at the point of retrieval." This allows the user to combine search terms with Boolean operators, functioning as a much more dynamic and unrestrained IR experience, enabling infinite combinations. Lois Chan points out that this method was not feasible in the libraries of yesteryear because users utilizing card catalogs were hindered from combining concepts.[17] Accordingly, pre-coordinated systems were understandably prevalent until the advent of the computer. In contemporary years systems such as MARC which use post-coordination often include subject headings with multiple simple one word descriptors. The user can combine these simple keywords during the search to come up with a wide array of records, although less specific to a complex subject area of inquiry.

According to a recent Library of Congress study and metaanalysis there is currently a rigorous debate over which method of subject representation will be used in the future across institutions, both public and private. The study found that pre-coordinated systems have strengths in: flexibility (easy to assemble and reassemble), library community support (other institutions prefer to work with them),

keywords (provides a proximity feature based upon subject area), clearer indication of general works, coherent ordering of subjects (topics have an internal structure and relationships, allowing for more coherent browsing), hierarchical displays (along with the former, improve browsability), a standard order which gives meaning ("The consistent use of a standard order to pre-coordinated strings in itself gives meaning to the words used") and notable relevance to search results (as things as manually sorted by a professional, relevant records are often returned in response to a query).[18] Disadvantages include: detrimental complexity (humans must devise elaborate systems to represent subject matter in this way, confusing both the end-user and the professional), too expensive (even when buttressed by cooperative efforts) and less flexible (because syntax must be considered).[19]

Ultimately while pre-coordination is superior for general browsing, it is anachronistic when used online, a remnant of a past era when catalog cards were a physical necessity to interface with knowledge. While the Library of Congress finally concluded in the study that they would continue to use pre-coordination, there was also a general impetus to streamline and re-envision the system so that it would not become completely useless in the web 2.0 age.

7. It is difficult to agree or to disagree with Morville's claim as it is logically fallacious: he fails to identify specifically who he is referring to or what their particular "prophecies" are. Artificial intelligence is not yet an endemic aspect of our information retrieval experience. The closest we come to encountering it in our everyday lives is the advanced heuristics of Google, which interprets the intent behind a poor search query and redirects the user to relevant content. Yet even this capacity is informed by the behavior of a larger end-user base and not in particularly by an intelligence which is judging the relevance of a query. If we substitute "artificial intelligence" with increased digitization, we may have

something to speak of, but it is not possible to discuss the implications of his claim verbatim, as to do so would be to follow false premises to inevitably false conclusions. For purposes of finishing this assignment I will consider IR technologies which attempt to inform the end-user (ranging from heuristics to user scripts) as synonymous with artificial intelligence.

I do not believe that making information systems more interactive and responsive will create anxiety or somehow unexpectedly roll back the progress we have made in IR so that it becomes slower and less effective. I can think of no historical precedent or possible extrapolation in which a technology's effectiveness advanced, degraded, and then caused sociopolitical dysfunctions. As long as the dubious "smart services" Morville ambiguously refers to do not become an exclusive means of accessing information, and there is no indication of this happening, then no such technological meltdown will occur. Smart services, effective heuristics and scripts informed by the user's tendencies, preferences and inclinations (a la Amazon's "you might also like these books!") help to bring technologically illiterate users into the digital age, and so enhance their lives with a wide range of new resources which up until recently only the technological experts have been privy to. Isn't it delightful when you type in something approximate and cryptic into a search engine and a list of popular alternatives becomes visible, often including what you meant to find? I would argue that this capacity is what is most important, rather than most dangerous, in our information systems. Advanced heuristics, the ability of an IR system to approximate what the user intended and then procedurally come up with suggestions, is perhaps the most critical determinant of quality.

Wow I just realized we are only supposed to pick 4 out of the 8 questions and want to cry now.

[1] G Salton, EA Fox, H Wu. "Extended Boolean information retrieval" 1983.

[2] Karen Markey, Pauline Atherton, Claudia Newton. "An analysis of controlled vocabulary and free text search statements in online searches" 1980.
[3] Kowalski, G; Maybury, M. "Information Storage and Retrieval Systems Theory and Implementation" Kluwer, Pp 97, 2000.
[4] Edmonds, P; Cotton, S. "SENSEVAL-2: Overview" In Proceedings of the Second International workshop on Evaluating Word Sense Disambiguation Systems. Toulouse, France, 2002. And Dervin, B., & Dewdney, P. (1986, Summer). Neutral questioning: A new approach to the reference interview. *RQ*, 506-513. Retrieved from http://slisweb.sjsu.edu/courses/restricted/dervin.pdf
[5] "Metadata." Oxford Digital Library. http://www.odl.ox.ac.uk/metadata.htm
[6] ibid
[7] Norm Medeiros. "A pioneering spirit: using administrative metadata to manage electronic resources." 2003.
[8] A Rajasekar, M Wan, R Moore, W Schroeder. "Metadata for multimedia documents." 2003.
[9] Richard Rubin. *Foundations of Library and Information Science*. 2nd edition. 2004. pp. 52.
[10] Geoffrey C. Bowker, Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences (Inside Technology).* 2000.
[11] Alex Thurman."Subject Indexing and Classification, 2002–2007" .http://www.ala.org/ala/mgrps/divs/alcts/resources/org/cat/research/subjindclass07.cfm
[12] Frank Brown. *Boolean Reasoning: The Logic of Boolean Equations.* 2003.
[13] Justin Blum. "Using AND, OR, and NOT (Boolean Operators)" Mathewson-IGT Knowledge Center. University of Nevada, Reno. http://www.library.unr.edu/instruction/help/booltips.html
[14] Library of Congress. "Library of Congress Subject Headings Pre- vs. Post-Coordination and Related Issues." http://www.loc.gov/catdir/cpso/pre_vs_post.pdf
[15] Lois Chan. *Cataloging and classification: an introduction.* 2007. pp. 204-206.
[16] Ibid.
[17] Ibid.
[18] Library of Congress pp. 7-8.
[19] Ibid 8.