Model Answers for Exam 2

Q1

For this question, consider the example of a search engine, such as Google – an organizing system for information resources on the Web. (Limit your analysis here to the search results themselves, not the advertisements that appear alongside them.)

Traditionally, information organization (IO) and information retrieval (IR) have been seen as separate disciplines. But in *The Discipline of Organizing*, we describe the process of computationally transforming resources for interactions as "re-organizing." Why? (1 point)

Do your interactions with a web search engine satisfy the Functional Requirements for Bibliographic Resources (finding, identifying, selecting, obtaining)? If so, how? If not, what is lacking? (4 points, 1 per functional requirement)

Since interactions involve "re-organizing," analyze your key interactions with a web search engine from the perspective of the five key design dimensions of organizing systems (what, who, when, why, and how much). (5 points - 1 per dimension)

Q1 Answer

Resources can be organized when they are being added to the organizing system, or being retrieved from the organizing system. In the case of a search engine, the process of computationally transforming resources for interactions is "re-organizing" because the set of query results for the specific query is "organized" at run time, i.e. after the user has run the query. When the web resources are being added to the collection of the search engine, they are processed, their metadata is created and then added to various collections. At run time, these meta data is used for further calculations (e.g. tf-idf) and then the query results are generated.

My interactions with a web search engine do satisfy the Functional Requirements for Bibliographic Resources, but in varying degrees:

- Finding - people use a search engine for different reasons - e.g. someone might use it to look up the price of his favorite book (in which case he knows exactly what he is looking for) or might be just querying iteratively to gain deeper understanding of a topic (in which case he might not even be aware of what exactly it is that he is searching for).
- Identifying - search results like Google make it quite easy to identify unique resources by removing possible duplicates as well as providing easy to read identifiers such as headings
- Selecting - based on the user's need, search engines allow their users to select a particular result and interact with it. But this selection is often implicit with the interactions that the user performs, rather than an explicit selection. E.g. a user might select a result by clicking on it to open the attached resource.
- Obtaining - since the search engine results are only resource descriptions, the actual resource has to be obtained by using an interaction such as clicking on it to open the resource itself.

Design dimensions of a web search engine:

1. What

All the possible documents that are available on the internet have to be organized. Even documents that the search engine finally decides to block or omit have to be tracked, so that they can be tracked.

2. Why

The resources are being organized so that users can find the relevant information on the internet quickly. Given the vast number of resources being added to the internet every hour, this is a much needed organizing system.

3. When

The organization happens when the resources are being added to the collection, when other resources are added to the collection as well as when the resources are being retrieved for a user's query.

Model Answers for Exam 2

4. Who

The organizing is being done by computational processes. These processes first crawl the web for new resources or updates to old ones. Then they are processed in various ways to create more metadata about them. Finally, when a user submits a query, the resources are evaluated for relevance and then reorganized according to their relevance scores before being presented to the user as query results.

5. How much is being organized

A lot. Not only is the scope of this is organizing system incredibly vast (probably the vastest in the world), there is a lot of additional information that needs to be created during the organizing itself. Any new resource that is added can possibly alter the rank and meta data of lots of other resources. E.g. incoming documents need to be scanned for links to pre-existing documents in the collection which would change their reputation. The incoming documents also have to be scanned for their content, and their frequencies calculated. When the user fires a query, the entire database needs to be searched and processed for the right relevant terms.

Q2

Resource descriptions can come from (1) personal experience, (2) the people who will interact with the resources, (3) experts in the resource domain, (4) automated or computational processes.

- Describe one strength and one weakness of creating resource descriptions by following each of these four methods (8 points, 1 per strength/weakness)
- How could you exploit the strength of the approaches and avoid their weaknesses by creating resource descriptions using multiple methods? Be specific. (2 points)

Q2 Answer 1

PERSONAL EXPERIENCE

Strength: Can potentially hold much more inner meaning, subtext for a user. Directly applicable with little translation.

Weakness:Those inner meanings aren't wholly transferrable.

PEOPLE WHO INTERACT WITH RESOURCES

Strength:Descriptions are probably conveyed in the vocabulary of the users.

Weakness:New users might not share the same vocabulary. And what happens if the organizing system is scaled to a new or expanded domain?

EXPERTS IN RESOURCE DOMAIN

Strength:Precision of the descriptions can be greatly enhanced to support more robust interactions with more granular and specific data.

Weakness:Experts outside the domain might not be able to appreciate the finer distinctions and data. What is gained in complexity may be lost in scalability.

AUTOMATED/COMPUTATIONAL PROCESSES

Strength:Discerning subtle differences and similarities through the parsing millions of resources could yield insights and distinguishing factors impossible for the human mind to conceive.

Weakness:The descriptions and their justifications might be completely impenetrable to human beings.

Essentially this is an exercise in translation and abstraction — by taking insights held by specific stakeholders and finding better ways to connect those insights with different stakeholders or a greater number of stakeholders.

For example, by pairing a computational process to apply descriptions to resources with experts in a domain, one might be able to ensure that the finer distinctions sussed out by a computer would be better understood or arbitrated by the people who know the resources best. Those domain experts indeed might be the only ones who could appreciate innovations or breakthroughs divined through a computational process. And they would be able to translate those discoveries into an organization system more easily understood by humans — at least human experts in that domain.

By pairing personal experience with the people who interact with the resource, again you're looking for ways to trade on commonalities between those stakeholders. For example, someone who works in a factory interacts with resources in her domain intimately. Perhaps she has a complex way of identifying, organizing, and utilizing those resources after 30 years of experience in the factory. Each factory worker might be in the same position. Now a new guy comes into the mix. Should he be expected to build up the same wealth of knowledge via trial and error interacting with those resources on his own? No. So a bridge must be built that can abstractify the wisdom of personal experience with the tactile experience of interacting with the resources at hand as a newcomer. Perhaps a sort of ad-hoc manual of information about the ins and outs of the factory could be shared among the veterans, thereby codifying some of those personal experiences into a vocabulary that would be more recognizable to the new user.

Q2 Answer 2

Personal Experience: A strength of creating resource descriptions from personal experience is that the descriptions will be meaningful and memorable for the person who created them and for anyone who shares enough of that person's experience. This would of course be relevant in a personal organizing system but may also be relevant in an organizing system for a close-knit group (e.g. family photo collection) or community of users who share some experience or viewpoint. An obvious weakness of resource descriptions based on personal experience is that the descriptions may not be meaningful to others and thus the organizing system could easily become obsolete or unusable (for example if resource descriptions of donors in a fundraising database are based too heavily on the personal experience of one or even a few staff members and those staff members all leave the organization the new staff may not be able to use the descriptions).

People who will interact with the resources: A strength of creating resource descriptions based on the people who will interact with the resources is that the interactions will be highly customized and applicable to the current users of the system. However, this approach is weak because there is significant potential for the descriptions to lose applicability when the group of people interacting with the resources changes or grows.

Experts in the resource domain: Creating resource descriptions based on experts in the resource domain would provide high precision for queries done by experts in the system and would allow the experts to select resources more easily. However, if the system is not designed for expert use alone non-expert users who are not familiar with the domain-specific terminology would have low recall because they would employ broader or less technical terms that may not appear in the resource descriptions.

Automated or computational processes: Automated or computational processes can avoid the issue of resource descriptions adhering too closely to the experience or knowledge of a specific person or group and are used in an attempt to create more "objective" resource descriptions. The strength of automated or computational descriptions is limited by the type of resources being described. For example, an automated description of quantifiable elements of a resource would be useful, things like the weight, dimensions, creation date, number of words, etc. However, an automated description that attempts to provide semantic meaning may not be useful because of the computer's inability to interpret effectively (e.g. an automated description of a photo is less likely to be useful than a description of a photo generated under any of the other three approaches).

Resource descriptions can support more and better interactions with a wider range of users by utilizing a combination of these approaches. This could be done by determining which resource descriptions are best generated by which methods. A resource description for a photograph could combine automatically generated metadata on the date and time created, shutter speed, aperture, and location and combine this with semantics from the other approaches. By mixing semantic descriptions from the other three approaches the a photograph organizing system could allow expert and non-expert users to achieve their desired interactions and could even accommodate descriptions based on personal experience for individual users or groups of users. Flickr gets close to this system because it incorporates automated data from the camera and allows users to enter resource descriptions that are designed for personal or group use (e.g. photos to be shared with family and friends) as well as allowing photographers to provide resource descriptions that are motivated by findability (for example, professional or semi-professional photographers who put their photos on Flickr so that they can be discovered and used by people outside of the photographers community). Flickr also allows for expert descriptions since users can enter any description they want but to get closer to taking advantage of expert resource descriptions Flickr would need to integrate an approach like that of Pandora by having experts in-house who would augment user-entered resource descriptions with descriptions based on controlled, domain-specific categories. That is, they could add descriptions specific to the domain of photography in general (using standards from museums) or specific to the various domains the photographs represent (e.g. scientific names for photos of flora and fauna).

Q3

Structural relationships can be directly perceived, unlike semantic ones, but we often identify or infer semantic relationships from structural ones.

- What are two common ways of using structural relationships inside of textbooks to convey semantic ones? Describe each relationship and its meaning. (2 points)
- Structural relationships in user interfaces can also reflect semantic ones. What are two common uses of structure in user interfaces to convey semantic information? (2 points)
- Presentation relationships sometimes convey or reinforce semantic distinctions but sometimes they don't. Give one example of a presentation rule in a style sheet or formatting guide that is semantic, and one that isn't. (2 points)
- If you are given a list of pairwise structural relationships (a connected to b, b connected to c, d connected to a, etc.), what techniques would you use to discover potentially important semantic relationships among them?
    - List 2 general techniques that work across different kinds of graphs/networks. (2 points)
    - List 2 techniques (distinct from those you noted above) that would help particularly with citation analysis. (2 points)

Q3 Answer

Textbooks have several ways to use structural relationships to imply semantic ones. The most obvious way to do so is within the table of contents. For example, having one section in large bold font, and having the following section in smaller font, with an indent, implies that the second section in some way semantically belongs the first one. This may work in a Biology textbook where the first heading reads '**Animals**' and the second heading reads '**Mammals'**.

Many textbooks use outlined sections, broken-off from the main progression of the text, to show case studies that are relevant to the material at hand. Having this structural separation, yet close proximity, implies that the content within that section does not fit within the general flow of the narrative, yet is still relevant to the material being discussed.

Similar methods are often used in interfaces to imply semantic value. Drop-down menus are a common method for this. Often times, similar values are clustered together, to make their use more intuitive. For example, when filling out an address form, possible country values will be contained together within a single drop-down menu. This tells the users that the items therein have equivalent semantic weight, yet only one value can be selected. Additionally, many digital interfaces use a header bar to indicate the primary interactions that they are able to support. Actions such as ‘open file’, ‘save file’, or ‘print’ are often clustered together in this manner.

Designation of font sizes is a presentation rule that often conveys semantics. Generally speaking, information in a larger font tends to be more important than information in a smaller font. Additionally, a bold font carries more weight than standard fonts. Displaying texts in italics can indicate that it is an aside or note, relative to the main body of text. Presenting data points on a map, as opposed to listing them out, can convey meaning about an item’s geographic relation to other items, without needing to explicitly write it out.

Weighting the edge within a graph, whether that is with an absolute number or with the thickness of the edge (in the presentation of a graph/network) would demonstrate the relative semantic importance of a relationship. A larger weight would give a relationship more value than one with a smaller weight. Additionally whether the relationship is a bi-directional or uni-directional (and if so, in which direction) one would imply something about the relationship. For example, two brothers would have a bi-directional relationship, yet a father and a son would have a uni-directional relationship (pointing from the father, to the son).

In regards to citation analysis, it is valuable to know if a single article (node) is cited by (has edges with) many other articles. This would imply that the article being cited has a large degree of impact in the academic community. If a given article is not cited often (has few relationships) it would imply that the article is either on a particularly niche topic, or does not have much academic impact. Additionally, you can use graph theory within citation analysis to uncover sub-graphs (distinct graphs contained within the larger graph) that would be indicative of ‘invisible colleges’. This may point to a group of academics studying one specific field that exists within a larger field. For example, articles on the asian longhorn beetle may be represented by a sub-graph in the larger academic graph for beetles.

Q4

In this semester’s reading, “How Psychiatry Went Crazy,” Carol Tavris reviews several books which discuss the problems with the Diagnostic and Statistic Manual of Mental Disorders (DSM), the authority by which psychiatrists diagnose mental disorders. One of these books concludes that the DSM, and the field of psychiatry in general, are efforts to turn the discontents of the human mind "into a catalog of suffering."

- What are three problems that the DSM faced in answering three of the organizing system questions you know from the case studies? (3 points)
- The first DSM in 1952 had 11 categories; the newest one has several hundred. In terms of organizing systems, what are at least two pros and two cons of increasing the categories of mental disorders? (4 points)
- Classification is principled, and maintaining the classification scheme over time must also follow principles. What are two challenges the DSM committee faces in maintaining their classification system over time? Explain what principles are at stake as they attempt to do this. (3 points)

Q4 Answer

The crucial overarching issue the DSM faces arises from the question of what is being organized. The DSM organizes mental health disorders, something which even experts in the field of psychiatry do not completely understand. As such, it is very difficult to determine what the resources are and how to describe them. This leads to the second issue faced by the DSM, who is doing the organizing, in a broad sense problems arise because the DSM is being organized by humans and humans do not have a full understanding of our own minds and emotions. Beyond this more philosophical issue there are other issues with who is doing the organizing because of the incentives for the DSM's organizers such as financial ties to pharmaceutical companies (who want more granular organization of diseases in order to create demand for more medication) and insurance claims (claims require a DSM diagnosis which could lead to incentives for more disorders to be classified so that it is easier for psychiatrists to bill). Additionally, there are time and data constraints on the organizers. A third question that raises issues for the DSM's organizers is how much to organize the resources. As noted there are incentives to create a more granular organizing system to satisfy pharmaceutical and insurance companies. In many cases disorders are identified and organized in order to prescribe medication for them. The increased granularity of the DSM also seems to be motivated by an attempt to be more scientific or to allow for and demonstrate greater understanding of human suffering. One of the authors mentioned in the article, Dr. Allen Frances, attempts to combat this conflation of progress in the field with more granular descriptions by pointing out that "It is impossible to define 'normal,' . . . let alone 'mental disorder.' But that doesn't mean . . . that we can't talk about the problems that cause human suffering."

One pro of increasing the categories of mental disorders is psychiatrists' increased ability to identify issues and treat patients suffering from those issues. Another pro may arise from the fact that identifying a disorder allows that particular disorder to be studied and understood more fully which may create better treatment and coping for those who suffer from the disorder. In some cases being able to find a name for something that has been plaguing you may give you some relief in itself in which case having a name for your specific issues could help you find a support community or encourage you to seek treatment when you may not have previously. There are also cons that arise from the increased categorization of mental disorders. A significant one, which is mentioned in the article, is the political nature of the identification of some disorders. While this can, and did, happen in a less granular DSM, it seems likely that when the threshold for categorizing a disorder is lowered more politically motivated disorder categories may be added. A second con, counter to the point above about potential positive effects of naming a disorder, is that more granular categorization of disorders subjects a broader range of the human experience to classification and as such pathologization. This can lead to over-medication and may also cause an individual, or the people around them, to over-identify with a disorder.

The DSM committee updates the DSM periodically to reflect changes in the field of psychiatry. This is an important maintenance activity because as the field of psychiatry progresses it gains a better (we hope!) understanding of mental health. Gary Greenberg highlights issues with data availability and deadlines that the committee faced in updating the DSM. A key principle of the DSM is that it is a scientific classification system based on empirical findings. However, because of issues with data availability, and particularly data readiness at the point of maintenance, this principle may be compromised in the name of expediency. Another principle of the DSM classification system is to base changes to the system on voting. While this principle is designed to draw on the knowledge of a broad group of experts it is vulnerable to, as Allen Frances says "subjective judgments that are inherently fallible and prey to capricious change." This is demonstrated by the fact that disorders have been removed from the DSM and then reinstated, such as narcissistic personality disorder.

Q5

Imagine you have a music collection with both digital and physical formats: mp3 files, CDs, vinyl albums, etc. Identify potential properties that you could use to organize the items in the collection (6 points, 1 point each):

- an intrinsic static property of individual physical resources
- a computed or calculated intrinsic static property of individual digital resources
- an extrinsic static property of individual resources
- an extrinsic dynamic property of individual resources
- an intrinsic static collection-level property
- an extrinsic dynamic collection-level property

Which of these properties would you use to organize your music collection? Using concepts from the course, explain why these properties are a good way to organize your personal collection. (2 points)

Model Answers for Exam 2

Now choose another mixed-format media collection, such as books (e-books and paper-based books), images (digital and prints), or something else. Which of the properties you listed for a music collection would **not** work well for this other domain? Why? (2 points)

Q5 Answer

- an intrinsic static property of individual physical resources
  - The publisher of the resource
- a computed or calculated intrinsic static property of individual digital resources
  - Time (length of play)
- an extrinsic static property of individual resources
  - The genre of the resource
- an extrinsic dynamic property of individual resources
  - Whether the resource is on a list of items to listen to this week
- an intrinsic static collection-level property
  - Total number of minutes of music that exists in the collection at a given time (as long as the collection exists in its current form, it can change owners or locations or whatever but the total play length will not change)
- an extrinsic dynamic collection-level property
  - Ranking (a number indicating the owner's favorites or values in the collection, ranked from most to least)

Of the properties I've listed, I would be most likely to use genre and author/artist to organize my music. Music to me is mood based , so when I'm interacting with my collection I will have a style of music in mind first (the why of organizing). Once I get to that part of my collection, I can't listen to all of it, I want to interact with one piece of music, so then I would "drill down" to a particular artist, then an album.

Some of the properties I've listed wouldn't work for books because they are specific to music (such as length of play, and total number of minutes on the collection). However, artist could easily become author, books also could be ranked as favorites or could be slated to be read that week, and books have genre as well. Length of play might be abstracted to just length here, and that would include text or number of words, too. One thing that might be unreasonable for books is a favorites list from week to week.

Q6

Given the following seven steps of text processing:

- Decoding
- Filtering
- Tokenization
- Normalization
- Stemming
- Stopwords
- Choosing index terms

Which of these steps are most impacted by the language of the text being processed? Why does the language matter? (3 points)

How are these steps impacted by the structure of the text? Compare and contrast this process for three unique types of documents (which contain text) on the document type spectrum. (5 points)

Which of these steps are most impacted by the IR model used (e.g. Boolean, TF/IDF Vector, or a more nuanced IR model of your choosing)? Provide examples of a step or multiple steps where one model is more effective than another model. (2 points)

Model Answers for Exam 2

Q6 Answer

Tokenization, normalization and stemming are most impacted. Tokenization is the process of separating sentences or words apart. The principle used to tokenizing one language doesn't always applied to other languages. It's possible we can use white spaces to tokenize most of words in English but it doesn't applied to Chinese because the words or phrases are not separated by white spaces. Normalization and stemming face the same problem as tokenization. Normalization means the symbols such as hyphens, accents, apostrophes or diacritics are removed and maybe also lowercase the words to make two words more identical. Take Spanish for example, putting the character n and ñ in a word makes the two words mean different things. Those words can never be semantically identical. Stemming removes the prefixes and suffixes of the word and leave the root form of the word. It's relatively easy to do stemming on English words compared to other languages. The morphological system for English is comparatively simpler than other languages. It more possible that we make more overstemming and understemming mistakes when we try to stem words in other languages.

Take plan-text articles, invoices and webpages for examples, there are decoding problems for all three of them. We don't know how to decode them if we don't know the encoding is ASCII, Unicode or other encoding methods. There is no markup in articles, so we don't have to perform filtering to them. However, the structure of the invoices and webpages are different from articles. They use specific structure to construct them. Invoices might be constructed by XML tags and webpages might be constructed by the mixup of XML and HTML tags. In this case, we have to filter the markups and try to retrieve the texts in them. Once we finish decoding and filtering, we can the pure texts each type of document. The rest of the steps are the same for all of them.

Choosing index terms is most impacted by the IR model used. Take boolean model for example, it has to manually choose and construct the index terms and their document frequency. In the other hand, the index terms are automatically selected in vector model. The process of choosing index terms in boolean model contains identifying different words, counting the number of each words in documents, combining the word-count results in different documents together. For vector model, the process only includes identifying different words and counting the number of each words in documents. There is one step missing because vector model using document vector to be the features when it performs comparison.

Q7

Some parts of the Web are more semantic than others. What is the motivation for making the web semantic? (2 points)

What parts are the least semantic and why? (2 points)

What parts are the most semantic and why? (2 points)

What are two current approaches for making the parts that are the least semantic more semantic? (2 points)

Some people think that libraries have great potential for making the web more semantic. How would this take place? (2 points)

Q7 Answer

The web started out for human use, but as it has grown the need for machine analysis of content has grown with it. Creating a more semantic web enables us to better search, process, and organize the vast amount of information that is out there. There is too much for humans to process alone.

The least semantic parts of the web are large unstructured blobs of text, old video formats that don't include markup of any kind, old sites that use HTML 1, images that don't include alt tags, and the like. Without semantic annotations or a clear structure, we have no clues as to relative importance or meaning of the content, which makes machine parsing of the information very difficult, but also this causes issues for those with disabilities (images without alt tags can't be interpreted by blind people, for example).

The most semantic parts of the web are those that are standardized/structured in a machine-readable format such as JSON, XML, or any other schema that can be easily interpreted across websites.

One approach for making the web more semantic is the use of XML to indicate semantic meaning. This is a good start, but it doesn't go far enough, because one site's <price> tag may refer to euros while another may refer to dollars (likewise a number indicated by <quantity> could mean bushels, dozens, pounds, kilograms, etc.). Another better approach is to use RDF (or RDFa), which is model that uses graphs in order to encode metadata and make statements machine processable. RDFa allows for a subject-predicate-object statement such as (student, plays, sport). A limitation with RDF is that it can't tell the difference between Bob Glushko and Dr. Robert J. Glushko, and it doesn't know that teaches is the inverse of taught by (OWL is another approach that tries to solve this by adding ontologies for semantic meaning).

Librarians and libraries have been organizing information for retrieval for a long time through authority control, classification, disambiguation, etc. – they make information discoverable! Librarians could be used to help transform the native data models of various data stores/institutions into RDF so that they are standardized and therefore interoperable/searchable by others outside that institution.