DCPAPERS

Leveraging Metadata Inductively and Subjectively

Stephen Davies, Roger King
University of Colorado, Boulder, Colorado, USA
{stephen.davies, rogerking}@colorado.edu

Abstract: Metadata has typically been used for deductive reasoning – applications take advantage of its precision to render objective conclusions about annotated sources that are 100% reliable. Metadata annotations could also be used inductively however, to allow systems to reason about the similarities that individual users perceive among instances. We explore how this notion could be of benefit to an information exploration agent.

Keywords: metadata; search engines; information retrieval; inductive reasoning; information space organization

1 Introduction

Metadata annotation can be seen as a way to make explicit the information that is only implicit (if not outright missing) in the annotated source itself. When a document is annotated with, for instance, a Dublin Core tag, agent programs no longer have to guess at its author, topics, or other attributes. They have reliable, unambiguous information at their disposal on which to base decisions. True, in any practical system the metadata is less than 100% accurate, but this is due to errors in the human's recording rather than to ambiguity in the representation. From a software agent's perspective, metadata is a gold mine of explicitly-recorded facts that require no guesswork to decipher

Traditionally, this has been used to enable *deductive* reasoning. Machine-readable data coupled with automated inference engines allow agents to draw powerful conclusions whose logical validity cannot be doubted. Hence the recent enthusiasm for ontology development, where the core axioms of a domain are codified into formal logic. The more domain knowledge that can be encoded into a precise representation, the broader the scope of the conclusions that a system can deduce. Such was the original vision of the Semantic Web, where more and more functions can be automated by a network of agents that truly "understand" the data they examine[2]

It occurs to us, however, that this metadata can be put to another use entirely: to facilitate *inductive* reasoning. Unlike deduction, which uses rigorous logic to derive irrefutable conclusions from base facts, induction reasons from particulars to generalities in a search for "mostly true" assertions. A deductive process would use "document X was authored by person Y" and "person Y belongs to institution Z" to derive "document X was authored by someone from institution Z." An inductive process, on the other hand, would use "document X_1 was authored by person Y" and "document X_2 was authored by person Y" to tentatively conclude "documents X_1 and X_2 are *similar* in some way." And if it turned out, for instance, that nearly all of the documents written by person Y had to do with topic T, then "person Y tends to write about topic T" would be another inductive conclusion.

The idea of induction – discovering general principles from a plethora of individual facts – has been studied for centuries, even before computing. Recently the twin fields of data mining and machine learning have produced an enormous literature on techniques for drawing just these kinds of conclusions. What is new about our work is that we propose to apply these principles to a realm that seems to have been largely overlooked; namely, metadata annotations. And we aim to do it in a *subjective* way, where each individual who approaches the data can express her own ideas about how it is related and how it should be organized.

This principle of subjectivity is fundamental to our work. The "meaning" of an entity (say, a document) is not absolute: it is subject to the interpretation of each user who perceives it. Metadata annotations are an attempt to record facts about an entity, only some of which may correlate with a user's perception. We envision an interface through which a user can express the semantics that she recognizes when she browses and interprets entities. The system, then, using fairly straightforward techniques, can inform the user of which metadata attributes may be significant indicators of whatever real-world phenomenon she has noticed. To take a simple example, a researcher may skim large amounts of literature in search

https://doi.org/10.23106/dcmi.952107840

skim large amounts of literature in search of papers relevant to a new idea she is considering. As she does this, she could place certain papers into categories according to which aspect of her idea they are pertinent to. Now by forming such groups and placing entities into them, she is implicitly making semantic judgments. She is expressing to the system that based on her intuitive analysis of their contents, the papers within a group are related to each other in some way.

But if metadata are associated with each of the papers, then it is a simple matter for the system to analyze them in search of commonalities within each group. This may be of tremendous benefit to the researcher. She may learn that the papers she has classified together in one of the groups tend to have a particular keyword or set of keywords of which she was previously unaware. Or it may turn out that although the papers in a group were written by different authors, many of those authors are affiliated with the same institution, or tend to publish in a particular journal, or frequently cite a key set of older publications. The system is using the metadata inductively not to draw any irrefutable conclusion, but simply to bring to the user's attention tendencies that may be useful to know about. This allows the "meaning" of the group to emerge, and its semantics generalized and extended to other, unseen data.

This is what we mean by leveraging metadata *subjectively*. We believe it is a mistake to try and form inductive hypotheses based solely on the objective annotations themselves, without taking into account the meanings that a human attaches to them. But when a system can combine the explicit metadata that describes entities with knowledge about how humans naturally group such entities together, the result can be a powerful method of forming and making use of knowledge.

2 Abstractness vs. concreteness

Humans excel at rendering intuitive judgments about what they perceive. We think at an abstract level, principally because the world we live in is so complex that we are forced to deal in generalizations to have any hope of coping with it. One of the great struggles throughout the history of computer science, in fact, has been the tension between the user's "abstractness" and the machine's "concreteness." Computers cannot tolerate vagueness or imprecision – they can only reason about particulars, and insist that everything they give their attention to be reduced to its most primitive elements. Thus users are forced to try and express their intentions in a form that does not do justice to their thoughts. For one example, consider information retrieval. When a user searches the Web, she does not usually have in mind the particular sequence of words that she will eventually discover in the form of a page. She begins rather with an abstract *idea*, a *concept*. The list of terms that she types into a search engine is merely an attempt to codify this into natural language elements as best she can.

We propose that a metadata-enabled search interface should permit a user to remain at the level of instances, rather than forcing her to try and guess at the lower-level fields that may reflect her abstract ideas. Instead, the *system* can do this for her. Given sufficient user input about what the instances "mean," and given rich enough metadata that describes each instance, a system can inductively infer which metadata fields support the user's subjective perception. Traditional information retrieval makes the assumption that the abstract notions data producers and consumers have in mind are reflected in the text itself. They presume that a searcher's mental conceptions can be characterized by sets of search terms, and that the "meaning" of a page can be discovered by mining for features (ie., words) that characterize it and embody its semantic notions. If this is true for natural language text – and the popularity of Web search engines certainly attests that it is to some degree – then shouldn't it be even more true for explicit metadata annotations? If the raw words on a page are a fairly reliable indicator of its meaning, and can be tapped inductively, how much more should this be the case for the annotations specifically *intended* to capture its meaning!

Note that we are addressing the general case here, not merely the case in which the user's mental concept happens to coincide directly with a metadata field. If the user simply wants to find documents by a particular author, for example, and the author is explicitly encoded in the metadata, then the "search" is straightforward. This is in fact what most metadata search engines support today. But when the user's concept is fuzzier – e.g., "I want

https://doi.org/10.23106/dcmi.952107840

fuzzier – e.g., "I want other documents like these three" – then an inductive process is necessary to nail down the user's idea in terms of the tangible features of the instances.

Humans reason about real-world entities, which are possessed of enormous complexity. Their perceptions are subtle, and may involve a myriad of characteristics. An office worker may comment about a new employee, "I don't know what it is, but Joe really reminds me of Bob somehow." Or a researcher may be overheard to say, "This paper's central idea reminds me of another paper I came across recently." When a user detects similarity between such entities, she is undoubtedly recognizing some real aspects that are shared between them. But a computer, of course, knows about only a tiny slice of the real-world object's complexity. Any idea, document, or person is boiled down to only a modest set of features for electronic storage. Now as long as these features somehow reflect the similarity that the user perceives, modern data mining techniques should be able to ferret out the subset of them that are trulyrelevant. But if the relevant aspects were not captured, the system is powerless to make any sense of it, no matter how sophisticated the algorithm.

In short, if we are to develop tools that allow users to query, explore, and reason based on abstract concepts, we need good underlying material for the system to work on. Inductive reasoning does no good if the semantics that users perceive in the real-world entities are unrecoverable or simply not present in the data stored about them. This demands richness, depth, and variety in metadata, to be sure. But it seems to us that in any case, the explicit metadata annotations are a far more reliable indicator of the meaning of an object than are the free-text words used to describe it.

3 A prototype for film exploration

We are beginning an investigation into the design of tools that leverage metadata inductively and subjectively. Our purpose is to raise the user's level of abstraction, so that he can think in the conceptual terms to which he is accustomed. The user will be able to ask abstract questions like "what other documents are similar to this one?" or "what is significant about this group of documents that I've identified?" or "in general, how does this group differ from that one?" Metadata is what the system will use to make such determinations.

Our initial prototype demonstrates these ideas on a subset of the Internet Movie Database (IMDb), an enormous digital library of films.[1] IMDb contains metadata annotations for the cast, crew, and production staff of nearly every movie ever produced. Also included are a set of keywords that describe interesting plot elements for each film. For instance, Alfred Hitchcock's "The 39 Steps" is annotated with such keywords as *based-on-novel*, *spy*, *chase*, *scotland*, *theatre*, and *conspiracy*. These keywords can be added by the public at large, and some popular movies have over a hundred of them.

We propose to take advantage of these annotations not only to improve searches for movies but also to help the user organize and understand the collection as a whole. In particular, we allow a user to work with *categories* that group together films that he perceives as similar in some way. He might, for instance, create a category called "violent war epics" and lump together a number of movies that fit that description. Not only would this maintain a list of specific films that he has included, but it would allow the system to learn from this list and then generalize – it could find other films he has not yet seen that might be probable candidates for the category. And in addition, he could ask more general questions at the category level, perhaps obtaining a list of the cinematographers, film editors, and special effects coordinators who commonly work on movies of this sub-genre.

All of this is based on inductive reasoning. "Violent war epics" is an abstract notion that resides only in the user's mind. It is not an explicit value for any metadata field that can be searched on. The user, in fact, would not even be able to define it precisely himself: he simply knows intuitively that certain movies fit this category, and others don't. Our supposition is that in many cases, the metadata contains information that is reflective of this, and that by examining the selected instances the system will be able to "lock on" to a set of reliable fields for classifying such films.

The application allows the user to create and name an abstract category. Then it shows him a series of titles one after the other. For each one, the user renders a semantic judgment, informing the system as to whether the movie is, or is not, in the category. The

https://doi.org/10.23106/dcmi.952107840

movie is, or is not, in the category The system thus retains a growing list of "examples" and "counterexamples" for the category and it uses these to select which title to show next. Hence it is able to achieve two objectives at once: it acquires data that allow it to reason inductively about what the category "means" in general, and it actively guides the exploration process towards films of interest. Each time the user gives positive feedback for a film, the system knows that it is on the right track with the instances that it is showing. Each time the user gives negative feedback, it assimilates the instance into its list of counterexamples, and self-corrects its course. In either case, it accumulates information that allows it to form a more precise understanding of the criteria for category membership. Our initial experiments with the system have been very promising. The keywords, particularly, seem to contain data that helps the system distinguish category members from nonmembers. Using the "violent war epics" example above, for instance, we were presented the following sequence of titles, each of which we judged to be a member or nonmember of the category:

Braveheart *(member)*
Monty Python and the Holy Grail *(nonmember)*
Millers Crossing *(nonmember)*
Return of the Jedi *(nonmember)*
Silence of the Lambs *(nonmember)*
The Apartment *(nonmember)*
The Return of the King *(member)*
Gladiator *(member)*
The Fellowship of the Ring *(member)*
The Two Towers *(member)*
Raiders of the Lost Ark *(nonmember)*
Stand by Me *(nonmember)*
Star Wars *(nonmember)*
Minority Report *(nonmember)*
Alien *(member)*
Aliens *(member)*
X2 *(nonmember)*
Terminator 2 *(member)*
The Empire Strikes Back *(nonmember)*
Platoon *(member)*
The Wild Bunch *(nonmember)*

These are the instances that the system suggested to us as we continued to give it formative feedback during a session. Its aim was to converge on the mental concept we had to converge on the mental concept we had in mind, and it obviously did: note that even some of the films we ultimately judged as *non*members began to approach our concept. For example, "Star Wars," "The Wild Bunch," and others are in fact epic action films, even if we did ultimately decide that they fell outside the scope of the category in question. In general, one can tell that the system is converging when it becomes more difficult for the user to decide if the instances are indeed members of the category The system is then effectively "exploring the edges" of the category in the semantic space. It is refining its understanding of exactly how the available metadata translates into membership or non-membership, giving it a more and more precise predictive capability. We can also ask the system for its "hypothesis" about what the category we had been defining meant. This command displays the metadata fields that the system finds to be statistically significant among the examples (and counterexamples) we had judged. We use a basic difference of means test to accomplish this, trimming all fields that do not meet a certain "Student's t" significance threshold. In our present case, the system gave the following output (trimmed slightly for brevity):

Hypothesis:

keyword=combat(3.35)
keyword=bravery(3.31)
keyword=blockbuster(2.93)
keyword=part-computer-animation(2.68)
keyword=decapitation(2.63)
keyword=heroism(2.57)
keyword=courage(2.57)
writer=WalshFrances(2.12)
director=JacksonPeterI (2.12)
NOT keyword=hologram(2.45)
NOT keyword=shootout(2.45)
NOT keyword=washington-d.c.(2.45)
NOT writer=LucasGeorge(2.45)
NOT keyword=cult-favorite(2.18)

The annotations preceded by "NOT" are indicative of counterexamples; the others are indicative of examples. The number in parentheses is the value of the significance statistic for that particular annotation. What this gives us is a composite look at how certain metadata tend to align with the abstract notion we have in mind. It is what the system has

https://doi.org/10.23106/dcmi.952107840

hypothesized is the "meaning" of our category, boiled down into the metadata attributes that tend to be associated with its members (or *not* associated with them, in the case of the "NOT" attributes.)

A system like this yields several advantages:

1. The user can examine this preliminary hypothesis for insight into what he might be unconsciously identifying with the category. This may be revealing in itself. In the above example, the user may never have realized that the type of film he has in mind often involves *computer-animation*. Or it may even cause him to reflect that perhaps it isn't so much the grand action sequences that really draw him to such films, but rather the more fundamental themes of *bravery* and *heroism* that they often involve. In general, a user may start out by suspecting that certain metadata fields will be adequate to capture his target concept, but he may see things in a new light when the system reports to him what is actually correlated. We envision the user being able to refine this hypothesis manually, removing attributes that are obviously spurious (e.g., the "washington-d.c." keyword in this case.) 2. The refined category can be used as a customized information retrieval construct. It is a detailed description of what kinds of things constitute a user's mental concept, and as we have seen, it can be used to quite reliably select other relevant instances.

3. Higher-level questions can be posed based on categories. After refining a definition, a user could ask, "what fraction of the entire film base is made up of 'violent war epics?'" or "who are the most common directors for such films?" or "how much longer (or shorter) do violent war epics tend to be than other films?" This is a powerful tool that allows a user to pose abstract questions about trends within a database. The system cannot with 100% reliability predict whether any given film is, or is not, a 'violent war epic' based on its metadata. But when the user has provided enough examples, we believe it will be able to answer such questions with a tolerable degree of accuracy. This will be a central focus of our future research.

4 Related work

Most metadata search engines (such as [3, 5, 8]) only permit the individual fields to be searched on. This is certainly useful, as its precision is a cut above that afforded by using only natural language techniques. But it requires that the user start out with definite ideas about which fields and which values are applicable to the search, rather than letting the system help determine that on his behalf. Inductive metadata analysis is not unknown [4], but it tends to be objective, attempting to find "the" best set of clusters within a data set, rather than involving a user's unique perceptions. Some work has been done on judging semantic similarity based on metadata, but this is used chiefly to rank items in a query's hit list [7], not to enable the kind of customized exploration we envision.

5 References

1. *The Internet Movie Database* 2004. Available at: http://www.imdb.com.
2. T. Berners-Lee, et al., "The Semantic Web," in *Scientific American*, May 2001.
3. J. Davies, et al. *QuizRDF: search technology for the semantic web* in *WWW2002 workshop on real world RDF & Semantic Web Applications, 11th International WWW Conference*. Hawaii, USA. 2002.
4. C. Fluit, et al., *Spectacle*, in *Towards the Semantic Web: Ontology-Driven Knowledge Management*, J. Davies, et al., Editors. 2003, John Wiley & Sons, Ltd.
5. J. Heflin and J. Hendler. *Searching the Web with SHOE*. in *Artificial Intelligence for Web Search. Papers from the AAAI Workshop*. Menlo Park, CA: AAAI/MIT Press. 2000.
6. D.L. Hicks and K. Tochtermann. *Personalizing information spaces: a metadata-based approach*. in *Proceedings of the International Conference on Dublin Core and Metadata Applications* Tokyo, Japan. 2001.
7. A. Maedche, et al. *SEAL - A framework for developing SEmantic portALs* in *Proceedings of the 18th British National Conference on Databases*. Oxford, UK: Springer-Verlag. 2001.
8. D. McGuinness, et al., *FindUR: Knowledge - enhanced online search*. 1998. Available at: http://www.research.att.com/~dlm/papers/findur-chi98.ps.

https://doi.org/10.23106/dcmi.952107840