a18 WEB SCIENCE

a18-01 - Presentación

Tarea: **Group 1. Read the following papers (or any other related one) and create a presentation of it.**

- A Manifesto for Web Science. Hall W, Hendler J, Staab S. http://www.webscience.org/manifesto/
- Hendler, Jim; Shadbolt, Nigel; Hall, Wendy; Berners-Lee, Tim; Weitzner, Daniel (2008). "Web science: an interdisciplinary approach to understanding the web". Communications of the ACM. 51 (7). doi:10.1145/1364782.1364798

- [ ] Buscar paper: Google Académico |
- [ ] Leer paper: https://eprints.soton.ac.uk/266555/1/CACM.pdf
  - [ ] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.490.2763&rep=rep1&type=pdf
  - [ ] https://cacm.acm.org/magazines/2008/7/5366-web-science/fulltext
  - [ ] Muchas alternativas y muchos formatos para el mismo paper.

https://cacm.acm.org/magazines/2008/7/5366-web-science/fulltext

Despite the web's great success as a technology and the significant amount of computing infrastructure on which it is built, it remains, as an entity, surprisingly unstudied. Here, we look at some of the technical and social challenges that must be overcome to model the Web as a whole, keep it growing, and understand its continuing social impact. A systems approach, in the sense of "systems biology," is needed if we are to be able to understand and engineer the future Web.

Despite the huge effect the Web has had on computing, as well as on the overall field of computer science, the best keyword indicator one can find in the ACM taxonomy, the one by which the field organizes many of its research papers and conferences, is "miscellaneous." Similarly, if you look at CS curricula in most universities worldwide you will find "Web design" is taught as a service course, along with, perhaps, a course on Web scripting languages. You are unlikely to find a course that teaches Web architecture or protocols. It is as if the Web, at least below the browser, simply does not exist. Many

"information schools" and "informatics departments" offer courses that focus on applications on the Web or on such topics as "Web 2.0," but the protocols, architectures, and underlying principles of the Web per se are rarely covered. Simplifying a bit, part of the reason for this is that networking has long been part of the systems curricula in many departments, and thus the Internet, defined via the TCP/IP networking protocols, has long been considered an important part of CS work. The Web, despite having its own protocols, algorithms, and architectural principles, is often viewed by people in the CS field as an application running on top of the Net, more than as an entity unto itself.

This is odd, as the Web is the most used and one of the most transformative applications in the history of computing, even of human communications. It has changed how those in academia teach, communicate, publish, and do research. In industry, it has not only created an entire sector (or, arguably, multiple sectors) but affected the communications and delivery of services across the entire industrial spectrum. In government, it has changed not only the nature of how governments communicate with their citizens but also how these populations communicate and even, in some cases, how they end up choosing their governments in the first place; recall the U.S. presidential debates in which candidates took questions online and through YouTube videos. It is estimated that the size of the human population is on the order of 1010 people, whereas the number of separate Web documents is more than 1011.

Computing has made significant contributions to the Web. Our everyday use of the Web depends on fundamental developments in CS that took place long before the Web was invented. Today's search engines are based on, for example, developments in information retrieval with a legacy going back to the 1960s. The innovations of the 1990s923 provide the crucial algorithms underlying modern search and are fundamental to Web use. New resources (such as Hadoop, lucene.apache.org/hadoop/, an open-source software framework that supports data-intensive distributed applications on large clusters of commodity computers) make it possible for students to explore these algorithms and experiment with large-scale Web-programming practices like MapReduce parallelism11 in a way not previously accessible beyond a few top universities. Other aspects of human interaction on the Web have been studied elsewhere. Of special note, many interesting aspects of the use of the Web (such as social networking, tagging, data integration, information retrieval, and Web ontologies) have become part of a new "social computing" area at some of the top information schools. They offer classes in the general properties of networks and interconnected systems in both the policy and political aspects of

computing and in the economics of computer use. However, in many of these courses, the Web itself is treated as a specific instantiation of more general principals. In other cases, the Web is treated primarily as a dynamic content mechanism that supports the social interactions among multiple browser users. Whether in CS studies or in information-school courses, the Web is often studied exclusively as the delivery vehicle for content, technical or social, rather than as an object of study in its own right.

Here, we present the emerging interdisciplinary field of Web science5,6 taking the Web as its primary object of study. We show there is significant interplay among the social interactions enabled by the Web's design, the scalable and open applications development mandated to support them, and the architectural and data requirements of these large-scale applications (see Figure 1). However, the study of the relationships among these levels is often hampered by the disciplinary boundaries that tend to separate the study of the underlying networking from the study of the social applications. We identify some of these relationships and briefly review the status of Web-related research within computing, We primarily focus on identifying emerging and extremely challenging problems researchers (in their role as Web scientists) need to explore.

What Is It?

Where physical science is commonly regarded as an analytic discipline that aims to find laws that generate or explain observed phenomena, CS is predominantly (though not exclusively) synthetic, in that formalisms and algorithms are created in order to support specific desired behaviors. Web science deliberately seeks to merge these two paradigms. The Web needs to be studied and understood as a phenomenon but also as something to be engineered for future growth and capabilities.

At the micro scale, the Web is an infrastructure of artificial languages and protocols; it is a piece of engineering. However, it is the interaction of human beings creating, linking, and consuming information that generates the Web's behavior as emergent properties at the macro scale. These properties often generate surprising properties that require new analytic methods to be understood. Some are desirable and therefore to be engineered in; others are undesirable and if possible engineered out. We also need to keep in mind that the Web is part of a wider system of human interaction; it has profoundly affected society, with each emerging wave creating new challenges and opportunities in making information more available to wider sectors of the population than ever before.

A large-scale system may have emergent properties not predictable by analyzing

micro technical and/or social effects.

It may seem that the best way to understand the Web is as a set of protocols that can be studied for their properties, with individual applications analyzed for their algorithmic properties. However, the Web wasn't (and still isn't) built using the specify, design, build, test development cycle CS has traditionally viewed as software engineering best practice.

Figure 2 outlines a new way of looking at Web development. A software application is designed based on an appropriate technology (such as algorithm and design) and with an envisioned "social" construct; it is indeed a contradiction in terms to talk about a Web application built for a single user on a single machine. The system is generally tested in a small group or deployed on a limited basis; the system's "micro" properties are thus tested. In some cases, when more and more people accept the micro system, accelerating "viral" scaling occurs. For example, when Mosaic, the first popular Web browser, was released publicly in 1992, the number of users quickly grew by several orders of magnitude, with more than a million downloads in the first year; for more recent examples, consider photo-sharing on Flickr, video-uploading on YouTube, and social-networking sites like mySpace and Facebook. The macro system, that is, the use of the micro system by many users interacting with one another in often-unpredicted ways, is far more interesting in and of itself and generally must be analyzed in ways that are different from the micro system. Also, these macro systems engender new challenges that do not occur at the micro scale; for example, the wide deployment of Mosaic led to a need for a way to find relevant material on the growing Web, and thus search became an important application, and later an industry, in its own right. In other cases, the large-scale system may have emergent properties that were not predictable by analyzing the micro technical and/or social effects. Dealing with these issues can lead to subsequent generations of technology. For example, the enormous success of search engines has inevitably yielded techniques to game the algorithms (an unexpected result) to improve search rank, leading, in turn, to the development of better search technologies to defeat the gaming.

The essence of our understanding of what succeeds on the Web and how to develop better Web applications is that we must create new ways to understand how to design systems to produce the effect we want. The best we can do today is design and build in the micro, hoping for the best, but how do we know if we've built in the right functionality to ensure the desired macroscale effects? How do we predict other side effects and the emergent properties of the macro?

Further, as the success or failure of a particular Web technology may involve aspects of social interaction among users, a topic we return to later, understanding the Web requires more than a simple analysis of technological issues but also of the social dynamic of perhaps millions of users.

Given the breadth of the Web and its inherently multi-user (social) nature, its science is necessarily interdisciplinary, involving at least mathematics, CS, artificial intelligence, sociology, psychology, biology, and economics. We invite computer scientists to expand the discipline by addressing the challenges following from the widespread adoption of the Web and its profound influence on social structures, political systems, commercial organizations, and educational institutions.

Beneath the Web Graph

One way to understand the Web, familiar to many in CS, is as a graph whose nodes are Web pages (defined as static HTML documents) and whose edges are the hypertext links among these nodes. This was named the "Web graph" in22, which also included the first related analysis. The in-degree of the Web graph was shown in Kleinberg et al.3 and Kumar et al.24 to follow a power-law distribution; a similar effect was shown in Broder et al.10 for the out-branching of vertices in the graph. An important result in Dill et al. 12 showed that large samples of the Web, generated through a variety of methods, all had similar properties—important as the Web graph grows, reported in 2005 to be on the order of seven million new pages a day.17 Various models have been proposed as to how the Web graph grows and which models best capture its evolution; see Donato et al.14 for an analysis of a number of these models and their properties.

Along with analyses of this graph and its growth, a number of algorithms have been devised to exploit various properties of the graph. For example, the HITS algorithmes and PageRank9 assume that the insertion of a hyperlink from one page to another can be taken as a sort of endorsement of the "authority" of the page being linked to, an assumption that led to the development of powerful search engines for finding pages on the Web. While modern search engines use a number of heuristics beyond these page-authority calculations, due in part to competitive pressure from those trying to spoof the algorithms and get a higher rank, these Web-graph-based models still form the heart of the critical crawlers and rank-assessment algorithms behind Web search.

The links in this Web graph represent single instantiations of the results of calling the HTTP protocol with a GET request that returns a particular representation (in this case an HTML page) of a document based on a universal resource identifier (URI) that serves as an identifier common across the entire

Web. So, for example, the URI http://www.acm.org/publications/cacm typed into a standard Web browser invokes the hypertext transfer protocol (HTTP) and returns an HTML page that contains content describing the publication Renown as Communications of the ACM. Note, however, that the content itself contains other URIs that are themselves pointers to objects that are also displayed (such as icons and images) and that the formatting of the page itself may require retrieving other resources (such as cascaded style sheets) or XML DTD documents. So what we might naively view as a single link from, say, a research group's Web page to an article on a Communications page will actually involve a number of requests among a number of servers; at the time of this writing, typing the URI for Communications into a browser will cause more than 20 different HTTP-GET requests to occur for seven different types of Web formats. Crawlers can capture these links and create the Web graph as, essentially, a static snapshot of the linking of the Web.

However, the Web graph is just one abstraction of the Web based on one part of the processing and protocols underlying its function. While it is an important result that the Web graph is scale-free, it is the design of the protocols and services that we now call the Web that makes it possible for it to be this way. The Web was built around a set of core design components defined in The Architecture of The World Wide Web, Volume 121 as "the identification of resources, the representation of resource state, and the protocols that support the interaction between agents and resources in the space."

A feature of the Web is that, depending on the details of a request, different representations may be served up to different requesters. For example, the HTML produced may vary based on conditions hidden from the client (such as which particular machines in a back-end server farm process the request) and by the server's customization of the response. Cookies, representing previous state, may also be used, causing different users to see different content (and thus have different links in the Web graph) based on earlier behavior and visits to the same or to other sites. This sort of user-dependent state is not directly accounted for in current Web-graph models.

There are also other ways the Web, as an application of the Internet, cannot simply be analyzed using the model of a quasi-static graph of linked hypertext pages. For example, many Web sites use Web forms to access a wealth of information behind the servers, where that information, sometimes called "the deep Web," is not visible in the Web model. For many sites, in which the applications's data forms a linked Web, the links are not explicit, and HTTP-POST requests are used instead of the HTTP-GETs in the Web graph. In other cases, these sites generate complex URIs that use GET requests to pass on

statea, thus obscuring the identity of the actual resources.
URIs that carry state are used heavily in Web applications but are, to date, largely unanalyzed. For example, in a June 2007 talk, Udi Manber, Google's VP of engineering, addressed the issue of why Web search is so difficult,25 explaining that on an average day, 20%–25% of the searches seen by Google have never been submitted before and that each of these searches generates a unique identifier (using server-specific encoding information). So a Web-graph model would represent only the requesting document (whether a user request or a request generated by, for example, a dynamic advertisement content request) linked to the www.google.com node. However if, as is widely reported, Google receives more than 100 million queries per day, and if 20% of them are unique, then more than 20 million links, represented as new URIs that encode the search term(s), should show up in the Web graph every day, or around 200 per second. Do these links follow the same power laws? Do the same growth models explain these behaviors? We simply don't know.
Analyzing the Web solely as a graph also ignores many of its dynamics (especially at short timescales). Many phenomena known to Web users (such as denial-of-service attacks caused by flooding a server and the need to click the same link multiple times before getting a response) cannot be explained by the Web-graph model and often can't be expressed in terms amenable to such graph-based analysis. Representing them at the networking level, ignoring protocols and how they work, also misses key aspects of the Web, as well as a number of behaviors that emerge from the interactions of millions of requests hitting many thousands of servers every second. Web dynamics were analyzed more than a decade ago,20 but the combination of (i) the exponential growth in the amount of Web content, (ii) the change in the number, power, and diversity of Web servers and applications, and (iii) the increasing number of diverse users from everywhere in the world makes a similar analysis impossible today without creating and validating new models of the Web's dynamics. Such models must also pay special attention to the details of the Web's architecture, as well as to the complexity of the interactions actually taking place there.

Today's interactive applications are very early social machines, limited by the fact that they are largely isolated one from another.

Additionally, modern, sophisticated Web sites provide powerful user-interface functionality by running large script systems within the browser. These applications access the underlying remote data model through Web APIs. This application architecture allows users and entrepreneurs to quickly build many

new forms of global systems using the processing power of users' machines and the storage capacity of a mass of conventional Web servers. Like the basic Web, each such system is interesting mainly for its emergent macro-scale properties, of which we have little understanding. Are such systems stable? Are they fair? Do they effectively create a new form of currency? And if they do should it be regulated?

Similarly, many user-generated content sites now store personal information yet have rather simplistic systems to restrict access to a person's "friends." This information is not available to wide-scale analysis. Some other sites must be allowed to access the sites by posing as the user or as a friend; a number of three-party authentication protocols are being deployed to allow this. A complex system is thus being built piece by piece, with no invariants (such as "my employer will never see this picture") assured for the user.

The purpose of this discussion is not to go into the detail of Web protocols or the relative merits of Web-modeling approaches but to stress that they are critical to the current and continued working of the Web. Understanding the protocols and issues is important to understanding the Web as a technical construct and to analyzing and modeling its dynamic nature. Our ability to engineer Web systems with desirable properties at scale requires that we understand these dynamics. This analysis and modeling are thus an important challenge to computer scientists if they are to be able to understand the growth and behaviors of the future Web, as well as to engineer systems with desired properties in a way that is significantly less hit or miss.

From Power Laws to People

Mathematically based analysis of the Web involves another potential failing. Whereas the structure and use of various Web sites (taken mathematically) may have interesting properties, these properties may not be very useful in explaining the behavior of the sites over time. Consider the following example: Wikipedia (www.wikipedia.org), the online wiki-based encyclopedia, includes more than two million articles in English and more than six million in all languages combined. They are hyperlinked, and it is logical to ask whether the hyperlinks have structure similar to those on the Web in general or whether, since this is a managed corpus, they have yet other properties.

Answering can be done in a number of ways; Figure 3 shows the result of one of them. In this case, DBPedia (dbpedia.org), which is a dump of the link structure of Wikipedia using the labeled links of the resource description framework, or RDF, has been analyzed with respect to the use of the link labels; that is, we are looking at the structure of Wikipedia as opposed to the linguistic content of its pages. The figure shows the same kind of Zipf-like

distribution found in the original Web graph analyses. There is also some evidence16and a lot of speculation29 that similar effects can be seen in the use of tags in Web-based tagging systems. Current research is also exploring whether these results depart from such models as preferential attachment3used to explain the scale-free features of Web graphs.

Unfortunately, whatever explains these effects, another aspect of Wikipedia's use is not explained by these models and does not necessarily follow from these properties. Wikipedia is built on top of the MediaWiki software package (www.mediawiki.org/wiki/MediaWiki), which is freely available and used in many other Web applications besides Wikipedia. While some of them have also been successful, many have failed to generate significant use. A purely "technological" explanation cannot account for this; rather, something about the organizational structures of Wikipedia and the needs of its users accounts for its success over other systems built from the same code base. The model by which articles are created, edited, and tracked is provided by the underlying technology. The social model enabled by humans interacting in ways allowed by that technology is more difficult to explain. The dynamics of any "social machine" are highly complex, and dozens of academic papers, from multiple disciplines, have been written about it; en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies uses Wikipedia itself to maintain an up-to-date reference list.

The idea of a social machine was introduced in Weaving the Web,8 which hypothesized that the architectural design of the Web would allow developers, and thus end users, to use computer technology to help provide the management function for social systems as they were realized online. The social machine includes the underlying technology (mediaWiki in the case of Wikipedia) but also the rules, policies, and organizational structures used to manage the technology. Examples abound on the Web today. Consider the coupling of the application design of blogging-support systems (such as LiveJournal and WordPress) with the social mechanisms provided by blogrolls, permalinks, and trackbacks that have led to the so-called blogosphere. Similarly, the protocols used by social networking sites like MySpace and Facebook have much in common, but the success or failure of the sites hinges on the rules, policies, and user communities they support. Given that the success or failure of Web technologies often seems to rely on these social features, the ability to engineer successful applications requires a better understanding of the features and functions of the social aspects of the systems.b

Today's interactive applications are very early social machines, limited by the fact that they are largely isolated from one another. We hypothesize that (i)

there are forms of social machine that will someday be significantly more effective than those we have today; (ii) that different social processes interlink in society and therefore must be interlinked on the Web; and (iii) that they are unlikely to be developed through a single deliberate effort in a single project or site; rather, technology is needed to allow user communities to construct, share, and adapt social machines so successful models evolve through trial, use, and refinement.

A number of research challenges and questions must be resolved before a new generation of interacting social machines can be created and evolved this way:

What are the fundamental theoretical properties of social machines, and what kinds of algorithms are needed to create them?;

What underlying architectural principles are needed to guide the design and efficient engineering of new Web infrastructure components for this social software?;

How can we extend the current Web infrastructure to provide mechanisms that make the social properties of information-sharing explicit and guarantee that the use of this information conforms to relevant social-policy expectations?; and

How do cultural differences affect the development and use of social mechanisms on the Web? As the Web is indeed worldwide, the properties desired by one culture may be seen as counterproductive by others. Can Web infrastructure help bridge cultural divides and/or increase cross-cultural understanding?

In addition, a crucial aspect of human interaction with information is our ability to represent and reason over such attributes as trustworthiness, reliability, and tacit expectations about the use of information, as well as about privacy, copyright, and other legal rules. While some of this information is available on the Web today, we lack structures for formally representing and computing over them. Traditional cryptographic security research and well-known access-control-policy frameworks have failed to meet these challenges in today's online environment and are thus insufficient as a foundation for the social machines of the future. Recent work on formal models for privacyb has demonstrated that traditional cryptographic approaches to privacy protection can fail in open Web environments. Similar problems with copyright enforcement have also hampered the flow of commercial and scholarly information on the Web.27 To this end, an exemplar Web science research area we are pursuing involves interdisciplinary research toward augmenting Web architecture with technical and social conventions that increase individual accountability to social and legal rules governing information use.31 Continued failure to develop scalable

models for handling policy will impede the ability of the Web to be the best possible medium for exchanging cultural, scientific, and political information.

The Web is changing at a rate that may be greater than even the most knowledgeable researcher's ability to observe it.

Further, we can see from the dramatic growth of new collaborative styles of creating and publishing information on the Web that many of the social institutions we rely on to judge trustworthiness and veracity are missing from our online information life. Being able to engineer the Web of the future requires not only understanding it as a computational structure but also how it interacts with and supports interaction among its users.

An important aspect of research exploring the influence of the Web on society involves online societies using Web infrastructure to support dynamic human interaction. This work—seen in trout.cpsr.org and other such efforts—explores how the Web can encourage more human engagement in the political sphere. Combining it with the emerging study of the Web and the coevolution of technology and social needs is an important focus of designing the future Web. 30

The Web of Data

This emerging area of study involves the heavy use of tagging provided by many of what are known as Web 2.0 technologies. Articles, blogs, photos, videos, and all manner of other Web resources may be annotated with usergenerated keywords, or tags, that can later be used for searching or browsing these resources. Much has been made of how "folksonomies," or taxonomies that emerge through the use of tags, can be used as metadata to help explain the content of the objects being described.

One aspect of tagging generating interest today is the need for "social context" in tagging.26 Many tags involve terms that are extremely ambiguous in a general context. For example, first names are popular tags on Flickr, though they are not good general search terms. On the other hand, in a specific social context (such as a particular person's photos), the same tag can be useful since it can designate a particular individual. The use of a tag as metadata often depends on such a context, and the "network effect" in these cites is thus socially organized.19

A more ambitious use of metadata involves recent applications of semantic Web technologies7 and represents an important paradigm shift that is a significant element of emerging Web technologies. The semantic Web represents a new level of abstraction from the underlying network infrastructure, as the Internet and

Web did earlier. The Internet allowed programmers to create programs that could communicate without concern for the network of cables through which the communication had to flow. The Web allows programmers and users to work with a set of interconnected documents without concern for the details of the computers storing and exchanging them.

The semantic Web will allow programmers and users alike to refer to real-world objects—people, chemicals, agreements, stars, whatever—without concern for the underlying documents in which these things, abstract and concrete, are described. While basic semantic Web technologies have been defined and are being deployed more widely, little work has sought to explain the effect of these new capabilities on the connections within the Web of people who use them.28

The semantic Web arena reflects two principle nexuses of activity. One tends to involve data (and the Web), and the other on the domain (and semantics). The first, based largely on innovation in data-integration applications, focuses on developing Web applications that employ only limited semantics but provide a powerful mechanism for linking data entities using the URIs that are the basis of the Web. Powered by the RDF, these applications focus largely on querying graph-oriented triple-store databases using the emerging SPARQL language, which helps create Web applications and portals that use REST-based models, integrating data from multiple sources without preexisting schema. The second, based largely on the Web Ontology Language, or OWL, looks to provide models that can be used to represent expressive semantic descriptions of application domains and provide inferencing power for both Web and non-Web applications that need a knowledge base.

Current research is exploring how the databases of the semantic Web relate to traditional database approaches and to scaling semantic Web stores to very large scales.1 In terms of modeling, one goal is to develop tools to speed inference in large knowledge bases (without sacrificing performance), including how to exploit trade-offs between expressivity and reasoning to provide the capabilities needed for Web scale.15 A market is beginning to emerge for "bottom-up" tools driven by data and "top-down" technologies driven by Web ontologies. Creating back-ends for the semantic Web is being transitioned (bottom-up) from an arcane art into an emerging Web application programming approach, as new open-source technologies integrate well with traditional Web servers. At the same time, new tools support ontology development and deployment (top-down), and tens of thousands of OWL ontologies are available for jumpstarting new domain-modeling efforts. In addition, approaches using rule-based reasoning modified for the Web have also gained attention.

4Engineering the future Web includes the design and use of these emerging technologies, along with how they differ from traditional approaches to databases, in one case creating back-ends for the semantic Web, in the other new tools for ontology-based applications.

The semantic Web is a key emerging technology on the Web, but, also, as we've discussed, there are different opinions as to what it is best for and, more important, what the macro effects might be. Our lack of a better understanding of how Web systems develop makes it difficult for us to know the kinds of effects the technology will produce at scale. What social consequences might there be from greater public exposure and the sharing of information hidden away in databases? A better understanding of how Web systems move from the micro to the macro scale would provide a better understanding of how they could be developed and what their potential societal effects might be.

Conclusion

The Web is different from most previously studied systems in that it is changing at a rate that may be of the same order as, or perhaps greater than, even the most knowledgeable researcher's ability to observe it. An unavoidable fact is that the future of human society is now inextricably linked to the future of the Web. We therefore have a duty to ensure that future Web development makes the world a better place. Corporations have a responsibility to ensure that the products and services they develop on the Web don't produce side effects that harm society, and governments and regulators have a responsibility to understand and anticipate the consequences of the laws and policies they enact and enforce.

We cannot achieve these aims until we better understand the complex, cross-disciplinary dynamics driving development on the Web—the main aim of Web science. Just as climate-change scientists have had to develop ways to gather and analyze evidence to prove or disprove theories about the effect of human behavior on the Earth's climate, Web scientists need new methodologies for gathering evidence and finding ways to anticipate how human behavior will affect development of a system that is evolving at such an amazing rate. We also must consider what would happen to society if access to the Web was denied to some or all and to raise awareness among major corporations and governments that the consequences of what appear to be relatively small decisions can profoundly affect society in the future by affecting Web development today. Computing plays a crucial role in the Web science vision, and much of what we know about the Web today is based on our understanding of it in a computational way. However, as we've explored here, significant research must still be done to be able to engineer future successful Web applications. We must understand

the Web as a dynamic and changing entity, exploring the emergent behaviors that arise from the "macro" interactions of people enabled by the Web's technology base. We must therefore understand the "social machines" that may be the critical difference between the success or failure of Web applications and learn to build them in a way that allows interlinking and sharing.

Acknowledgments

Figure 2 is taken from talks Tim Berners-Lee gave in 2007 (www.w3.org/2007/Talks/1018-websci-mit-tbl/Overview.html). We also thank the other members of the WSRI Scientific Council (webscience.org/about/people/) for input relating to the goals of Web science and the interaction of the Web and computer and information sciences. We are indebted to Konstantin Mertsalov of Rensselaer Polytechnic Institute for the DBpedia analysis discussed in the section on power laws.

References

1. Abadi, D., Marcus, A., Madden, S., and Hollenbach, K. Scalable semantic Web data management using vertical partitioning. In Proceedings of the 33rd International Conference on Very Large Data Bases (Vienna, Austria, Sept. 23–27). VLDB Endowment, Heidelberg, 2007.

2. Backstron, L., Dwork, C., and Kleinberg, J. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th International World Wide Web Conference (Banff, Alberta, Canada, May 8–12). ACM Press, New York, 2007.

3. Barabasi, A. and Albert, A. Emergence of scaling in random networks. Science 286 (1999).

4. Berners-Lee, T., Connolly, D., Kagal, L., Scharf, Y., and Hendler, J. N3Logic: A logical framework for the World Wide Web. Theory and Practice of Logic Programming (2008).

5. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., and Wietzner, D. Creating a science of the Web. Science 311 (2006).

6. Berners-Lee, T., Hall, W., Hendler, J., O'Hara, K., Shadbolt, N., and Weitzner, D. A framework for Web science. Foundations and Trends in Web Science 1, 1 (Sept. 2006).

7. Berners-Lee, T., Hendler, J., and Lassila, O. The semantic Web. Scientific American (May 2001).

8. Berners-Lee, T. and Fischetti, M. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Harper Collins, New York, 1999.

9. Brin, S. and Page, L. The anatomy of large-scale hypertextural Web search engine. Presented at the Sixth International World Wide Web Conference (Santa Clara, CA, Apr. 7–11, 1997).

10. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the Web. In Proceedings of the Ninth International World Wide Web Conference(Amsterdam, The Netherlands, May 15–19). Elsevier, Amsterdam, The Netherlands, 2000.
11. Dean, J. and Ghemawat, S. MapReduce: Simplified data processing on large clusters. In Proceedings of the Sixth Symposium on Operating System Design and Implementation (San Francisco, Dec. 6–8). USENIX Association, Berkeley, CA, 2004.
12. Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D. and Tomkins, A. Self-similarity in the Web. In Proceedings of the 27th International Conference on Very Large Data Bases (Rome, Italy, Sept. 11–14). Morgan Kaufmann Publishers, Inc., San Francisco, 2001.
13. Domingos, P., Golbeck, J., Mika, P., and Nowak, A. Social networks and intelligent systems. IEEE Intelligent Systems, Trends & Controversies 20, 1 (Jan./Feb. 2005).
14. Donato, D., Laura, L., Leonardi, S., and Millozzi, S. The Web as a graph: How far we are. ACM Transactions on Internet Technology 7, 1 (Feb. 2007).
15. Fokoue, A., Kershenbaum, A., Ma, L., Schonberg, E., and Srinivas, K. The Summary Abox: Cutting ontologies down to size. In Proceedings of the International Semantic Web Conference (Athens, GA, Nov. 5–9). Springer Berlin, Heidelberg, 2006.
16. Golder, S. and Huberman, B. The Structure of Collaborative Tagging Systems (2005); arxiv.org/abs/cs/0508082.
17. Gulli, A. and Signorini, A. The indexable Web is more than 11.5 billion pages. In the special-interest tracks and posters of the 14th International World Wide Web Conference (Chiba, Japan, May 10–14). ACM Press, New York, 2005.
18. Hendler, J. Web 3.0: Semantic Web chicken farms. IEEE Computer 41, 1 (Jan. 2008).
19. Hendler, J. and Golbeck, J. Metcalfe's Law, Web 2.0, and the semantic Web. Journal of Web Semantics 6,1 (Feb. 2008).
20. Huberman, B. and Lukose, R. Social dilemmas and Internet congestion. Science 277, 5325 (July 1997).
21. Jacobs, I. and Walsh, N. Architecture of the World Wide Web, Vol. One. W3C Recommendation, Dec. 15, 2004; www.w3.org/TR/webarch/.
22. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. The Web as a graph: Measurements, models, and methods. In Proceedings of the Fifth Annual International Conference on Computing and Combinatorics (Tokyo, July 26–28). Springer, New York, 1999.
23. Kleinberg, J. Authoritative sources in a hyperlinked environment. Journal

of the ACM 46, 5 (Sept. 1997).
24. Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. Trawling the Web for emerging cyber communities. In Proceedings of the Eighth International World Wide Web Conference (Toronto, May 11–14). Elsevier North-Holland, Inc., New York, 1999.
25. Manber, U. Why Search Is a Hard Problem. Presentation at Supernova 2007 (San Francisco, June 16–18, 2008); www.readwriteweb.com/archives/udi_manber_search_is_a_hard_problem.php
26. Marcus, A. and Perez, A. m-YouTube mobile UI: Video selection based on social influence. In Proceedings of the 12th International HCI Conference (Beijing, July 22–27). Springer, 2007.
27. Samuelson, P. Copyright's fair use doctrine and digital data. Commun. ACM 37, 1 (Jan. 1994), 21–27.
28. Shadbolt, N., Hall, W., and Berners-Lee, T. The semantic Web revisited. IEEE Intelligent Systems 21, 3 (May/June 2006).
29. Shirky, C. Power Laws, Weblogs, and Inequality In Clay Shirky's blog (2003); www.shirky.com/writings/powerlaw_weblog.html.
30. Shneiderman, B. Web science: A provocative invitation to computer science. Commun. ACM 50, 6 (June 2007), 25–27.
31. Weitzner, D., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., and Sussman, G. Information accountability. Commun. ACM 51, 6 (June 2008).
32. Weitzner, D., Hendler, J., Berners-Lee., T., and Connolly, D. Creating a policy-aware Web: Discretionary, rule-based access for the World Wide Web. In Web and Information Security, E. Ferrari and B. Thuraisingham, Eds. IRM Press, Hershey, PA, 2006.

Authors

James Hendler (hendler@cs.rpi.edu) is the Tetherless World Chair of Computer and Cognitive Science at Rensselaer Polytechnic Institute, Troy, NY.
Nigel Shadbolt (nrs@ecs.soton.ac.uk) is professor of artificial intelligence and deputy head of the School of Electronics and Computer Science at Southampton University, Southampton, U.K.
Wendy Hall (wh@ecs.soton.ac.uk) is a professor of computer science at the University of Southampton, Southampton, U.K.
Tim Berners-Lee (timbl@csail.mit.edu) is the Director of the World Wide Web Consortium and holds the 3Com Founders chair and is a senior research scientist in the Laboratory for Computer Science and Artificial Intelligence at the Massachusetts Institute of Technology, Cambridge, MA.
Daniel Weitzner (djweitzner@csail.mit.edu) is director of the Massachusetts Institute of Technology Decentralized Information Group and principle research

scientist in the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA.

Footnotes

a. These characters, including ?.#, =, and &, followed by keywords, may follow the last "slash" in the URI, thus making for the long URIs often generated by dynamic content servers.

b. When we say "success" or "failure," we are referring not to the business factors that determine whether, for example, Facebook or MySpace will attract more users but to the success or failure of the sites to provide the particular types of social interaction for which they are designed.

Funding for this work comes from the U.S. National Science Foundation (Policy Aware Web and Transparency Aware Data Mining Projects), iARPA (End-to-End Semantic Accountability), the U.K. Engineering and Physical Sciences Research Council (Advanced Knowledge Technologies Project), and the U.S. Army Research Laboratory and U.K. Ministry of Defence (U.S./U.K. Information Technology Alliance). We also thank industrial and individual donors to the authors' research at RPI, Southampton, and MIT and to the Web Science Research Initiative (www.webscience.org).

DOI: http://doi.acm.org/10.1145/1364782.1364798

Figures

Figure 1. The social interactions enabled by the Web put demands on the Web applications behind them, in turn putting further demands on the Web's infrastructure.

Figure 2. The Web presents new challenges to software engineering and application development.

Figure 3. Results of an analysis of the link structure of Wikipedia with respect to the use of link labels, not the linguistic content of pages.

a18-02 Introducción y presentaciones

Oscar Corcho

3 presentaciones hoy,

Nosotros la próxima semana

visión general de la disciplina de web science

Reflexionar sobre los elementos fundamentales que estamos analizando

Web, características principales

Perspectiva técnica

Identificadores VS Contenidos

Qué medidas consideramos para evaluar la web

- Tratar la web como una red, un grafo muy interconectada de elemntos
  - al principio era solo **documentos** html
  - ahora hay otros recursos, vídeos
- Cuando yo miro la web desde el lado:
  - Físico ➜ Servidores
  - Documental ➜ enlaces
  - Redes sociales ➜

¿Qué medidas utilizamos?

Teoría de grafos y de redes , super importante para webscientist
número de nodos, arcos, diámetro y radio, densidad de la red, grado de entrada salida en cada nodo (nodos prioritarios, influencers de instagram), clusterings, rutas , componentes conectados fuerte y débil, centrality, búsqueda de nodos clave, y búsqueda de comunidades

Las GAFA == Google , Amazon, Facebook y Apple
Pueden tener
Están centralizando
Si eres un científico que no trabajas en una GAFA, no tienes los mismos recursos. No tienes acceso a datos
amenaza a gafa: O nos dais las fuentes de datos , o no pertenecéis a

Carlos en topic models

- Internet Archive
- [SW Heritage](Software Heritage)
- Common Crawl
  - Hace el mismo trabajo que GAFA para hacer un crawl de la web
- Wikipedia/DBpedia (de los infoboxes , creative commons, puedes hacer lo que quieras)
  - se utiliza para
- WikiData
  - eran los de wikipedia
  - ahora están en google research
  - que la gente pueda subir datos
    - datos sobre calles
- OpenStreetMap
  - Alternativa a google maps, originalmente alternativa a ordinal surbay, cobraba por cada dato
  - política de datos abiertas
  - Que estén disponibles en la web, no quiere decir que es abierto, no es legal
    - Geo hackers
    - farolas para contaminación lumínica, grado de luz azul
  - Geoname
    - menos potente que OSM
  - Twitter
    - no puedes almacenar el texto, solo los links
    - User generated content ➜ © Twitter
  - Knowledge graph de varias compañías
    - Search API ➜
      - podemos buscar entidades, con una query
      - nos devuelve con schema.org
      - JSONLD
      - @id ➜ kg:/m/0d1567
    - Agregadores únicos de contenido, porque partir de ahí pueden ofrecer cualquier
  - Guerra abierta entre amazon y google , sobre las direcciones postales
    - En google places
    - Abajo hay una dirección rara: MCQ2+V5 El Molar
      - Si amazon utiliza estos códigos, tiene que pagar a google
      - El código se los das a amazon y en ese código está asociado a las entregas
    -
- Geonames
- Twitter

- Knowledge graphs from several companies

Característica de cualquier análisis

Desde el punto de vista computacional

- Small world
  - El mundo es pequeño, los pasos para conectar 2 nodos es pequeño
  - Cada vez será más pequeño
  - El número de nodos crece, pero los pasos de un nodo a otro disminuye
  - 6 grados de de separación. Erdos number
- Scale free
  - Hubs hiperconectadoá que cada vez se conectan aún más
    - Ejemplo: DBPedia in Linked Data

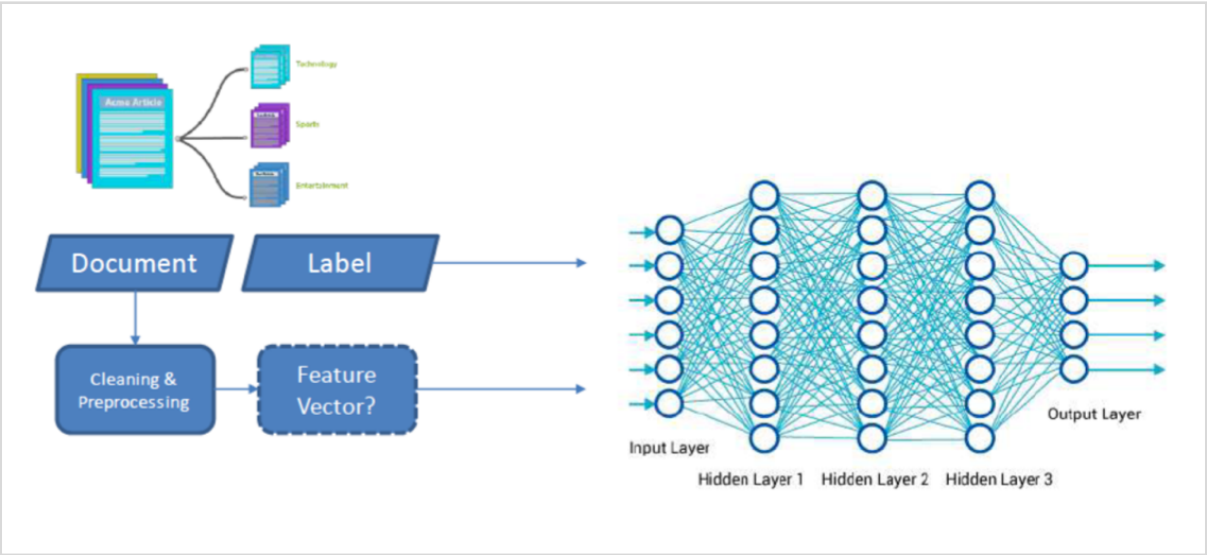
Tópicos que vamos a ver

1. Understanding the web (With NLP glasses)
   1. Web de documentos
      1. Escala de documentos muy grande. A gran escala
      2. Grandes corpus documentales, con los que estaremos trabajando
      3. Computacionalmente tendremos
   2. Cómo modelados estos corpus de texto
      1. Intentamos tratar descripciones más corta que represente de manera eficiente
      2. Buscamos el procedimiento eficiente de las largas colecciones
         1. Hacemos resúmenes de las colecciones
         2. para poder procesarlas computacionalmente
   3. Preservaremos las relaciones estadísticas para : clasificación, novelty, detection,

Como solemos trabajar con los corpus documentales

Conjunto de documentos

- Cada documento, puede estar etiquetado
  - con noticias, nos dirá la categoría, los
- Cleaning&preprocessing ➜ vectores de características ➜ los tratamos de cualquier manera



- Vector space Model - Columnas de Términos - lemas (muy dependiente del idioma)
  - Bolsas de palabras
  - Representamos cada párrafo en un vector de muy grandes dimensiones
  - tabla de frecuencia de palabras
    - no hay contexto
    - hay palabras `sparse` que no aparecen y pondremos muchos 0's
      - *Latent Semantic Analysis* : Técnicas para reducir la dimensionalidad: Buscas autovectores, y una matriz , la conviertes en multiplicación de varias matrices
    - la tabla es gigante
-
- LSA/LSI, Latent Semantic Analysis/Indexing: No puedo explicar fácilmente los resultados
- Latent Dirichlet Allocation (LDA)
  - Tópicos, conjuntos de palabras, distribuciones
  - Decidimos usar 100 tópicos, reducimos la dimensionalidad de las columnas a 100, y cada término le damos una probabilidad a cada tópico.
    - Cada tópico tiene una lista de términos
  - Reducimos nuestro vocabulario de Términos a tópicos
  -
  - De documentos a tópicos

- Word2Vec : cada palabra no es una bolsa de palabras , está asociada con las 5 siguientes y las 5 anteriores
  - Palabras a embeddings
  - Permite hacer aproximaciones de que palabras son similares , aparecen en espacios multidimensionales
  - Reglas de 3 de estructuras de inferencia, es posible hacerlos con embeddings
  - Documentos a vectores

> Cálculo del coseno entre dos vectores para comparar

Reconocimiento de entidades

Panama papers, Probablistic topic models

Segundo bloque: Social machines

Cómo las entidades físicas aparecen en la web ofrecen comportamientos sociales
Self-organization, self-adaptation y self-maintenance
Desarrollos de sistemas que sean autónomos

Evoluciones de las comunidades
Se usa mucho en los robots, coches autónomos. Serían comunidades. Rutas mejores con comunidades.

Perspectiva biológica, como se unen entre ellos para ver la relación

Natural System's Societies

Hacemos acuerdos, negociamos, comunicaciones
Somos self-organizing , sin preparar una reunión sabemos cuales son nuestros roles
Lo veremos en lo micro y en lo macro

Social Computing: Social Behavior JOIN Computational system

- Representing social Information and Knowledge
- Agent-based social modelling
- Human-agent interacion
- Analysis and prediction
- Negotiation and decision making systems
    - Como conseguir que estén felices

Ciencia ciudadana

Scientific research: Con científicos profesionales y ciudadanos
Zoonivers
NightKnights.eu

PRESENTACIONES

Objetivos de las presentaciones

entrar en faena, hacer análisis, y dar feedback sobre cómo hacer el estado del rte.
Feedback de cómo presentar
Presentaciones orientadas a la investigación.

Presentación 1 - tema 3

Grupo 3: Current WS topics
Luciano García Giordano
Salvador González Gerpe
11th presentación

Mejoras en presentaciones

- Explicar muy bien la metodología que hemos seguido, en el análisis del estado de arte
  - Qué habéis leído realmente
  - Método de adquisición de los papers
  - Me he leído todos los procedimos
  - Keywork research, he recuperado 400 documentos y he filtrado los que tienen esta palabra adicional.
  - Permite saber si habéis tenido algún fallo en tu metodología
    - qué pasos hemos seguido
  - Keynote speeches , posters
    - penúltimo era ing. ontologica , OKU, 20 años con un mundo feliz
  - Análisis de metodología
    - No describimos Methods & Tools
- Siguiente paso de metodología
  - Tópicos,
  - Son los keywords del paper? o los habéis creado vosotros?
    - buscar keywords habituales , tesauros y analizar
      - lo que hace un bibliotecario
  - Qué materiales habéis generado intermedios
  - Excel, papeles,
  - Para asegurar la reproducibilidad de estos experimentos
    - o que puedan ofrecer sugerencias
- Presentación de tópicos

    - presentación butterfly
    - tamaño del círculo afecta
    - justificación , **contar alternativas de presentar**
        - nubes de palabras
    - diferencia entre colores en la leyenda
        - menos de 3 papers , en azul
- Hay conflicto entre los tópicos. De la propia comunidad dice a los que estáis de fuera
    - keywords de los artículos y en qué tracks está
    - análisis de
- Open Science y digital science
    - Se han creado conferencias específicas hace 2 años
    - Blockchain y cryptology ➜ muy técnica
- Técnicas
    - sería interesante modificar los tamaños de los círculos para ver qué técnicas son más usadas
    - qué datasets y software se han generado en la conferencia

Presentación 2 - Tópico 1

Daniel Bermejo y Federico

- Qué es la Web Science
- Por qué es necesario?
    - la web cambia el mundo y viceversa
- La web con el mundo real
- La web - con la IA
    - interactuamos entre humanos y los robots
- Ambivalencias de la web
    - con doble sentido
    - la web se puede usar para bien o para mal
    - libertad y calidad de la información
        - fake news
    - personalización vs privacidad
    - Manipulación de las masas
    - alcance de la web
        - que la web llegue a todo el mundo
        - FB quiere que llegue todo el mundo, habría un monopolio sobre la web
    - Crecimiento y sostenibilidad
        - gastos de infraestructuras
- Web linking all forms of intelligence

- Retos de AI y la web
  - individual - grupal
    - individual , uso de la info, agentes de volumen de información
- Como gobernar la web
- Micro to macro
  - como crecen las aplicaciones
- Conclusión
  - mantener la web creciendo
  - analizar las implicaciones sociales y técnicas
- Referencias
  - manifesto
  - artículo

Feedback

añadir el número de las páginas de diapositivas

Son muy bla bla, más filosóficos, de visión.
Difícil es condensar exponer todos los temas de la manera
Tres de ellos son gente muy potente.
Y hacer una presentación que expone nuevas perspectivas

En estado de arte, metodología
En qué te basas
METODOLOGÍA
Referencia al principio del todo.

Corpus documental (INTRODUCCIÓN)
Hoy lo que vamos a contaros, está basado en esta documentación , en esta literatura. Y además de este

Hacer una ted talk

Añadir Referencias
From Micro to Macro, referenciar el párrafo
entrecomillo el párrafo “”
Wikipedia , un blog, Fuente gris, es

Técnica útil

Añadir testimonials

Meter Quotes del artículo

Incluir un paraguas de todos los tópicos que voy a tratar
Dibujar un camino del journey
No ir de tópico a tópico
Si ahora tengo el Gorro de computer science, si tengo
Multiples perspectivas. Distintos stakeholders

- Legal
- Sociologica
- Computacional
- Biológica

De dónde venimos, Dónde estoy, y los challenges
W questions, When, What, How

Slide de AI-World , para que la cognitiva no se pierda, ir añadiendo texto.
Un color distinto los mensajes
Términos que de individual y grupal divididos que vayan apareciendo

Usar las slides para que la gente no se pierda.
Bullet points, para take home messages

Síntesis,
Reflexión

Ideas

sacar la frecuencia de repetición de las palabras del artículo
¿Quién de vosotros usa la web? ¿Quién de vosotros realmente conoce la web?
Juego kazhoo?
vídeo micelios

Recomendation systems

Enviar feedback

a18-03 - Recommendation Systems

What is the problem

Retos desde el punto de vista de web science

- Evaluar la calidad
  - felicidad
- Cómo sabemos que no
- Qué es web-related recommendation for you
  - no hay mucha diferencia
  - antes recomendaban las agencias de viaje
- Las recomendaciones ha surgido especialmente en e-commerce

Buscar items o servicios que el usuario le gusta o quiere

> Estamos haciendo que los usuarios quieran algo.

Dos metas muy claras

- amazon
  - libros y películas
  - correlación, tenían una matriz de usuarios - items
    - muy exitoso , incluso con básicas técnicas estadísticas
- netflix
  - netflix challenge
    - Muy importante para entender que la recomendación es un negocio donde se puede hacer una investigación
    - 1M$ para el mejor recomendado de películas
    - se comprobaba con el gold star
  - problema
    - cambiamos la intención del usuario
    -

> cómo evaluar un sistema de recomendación

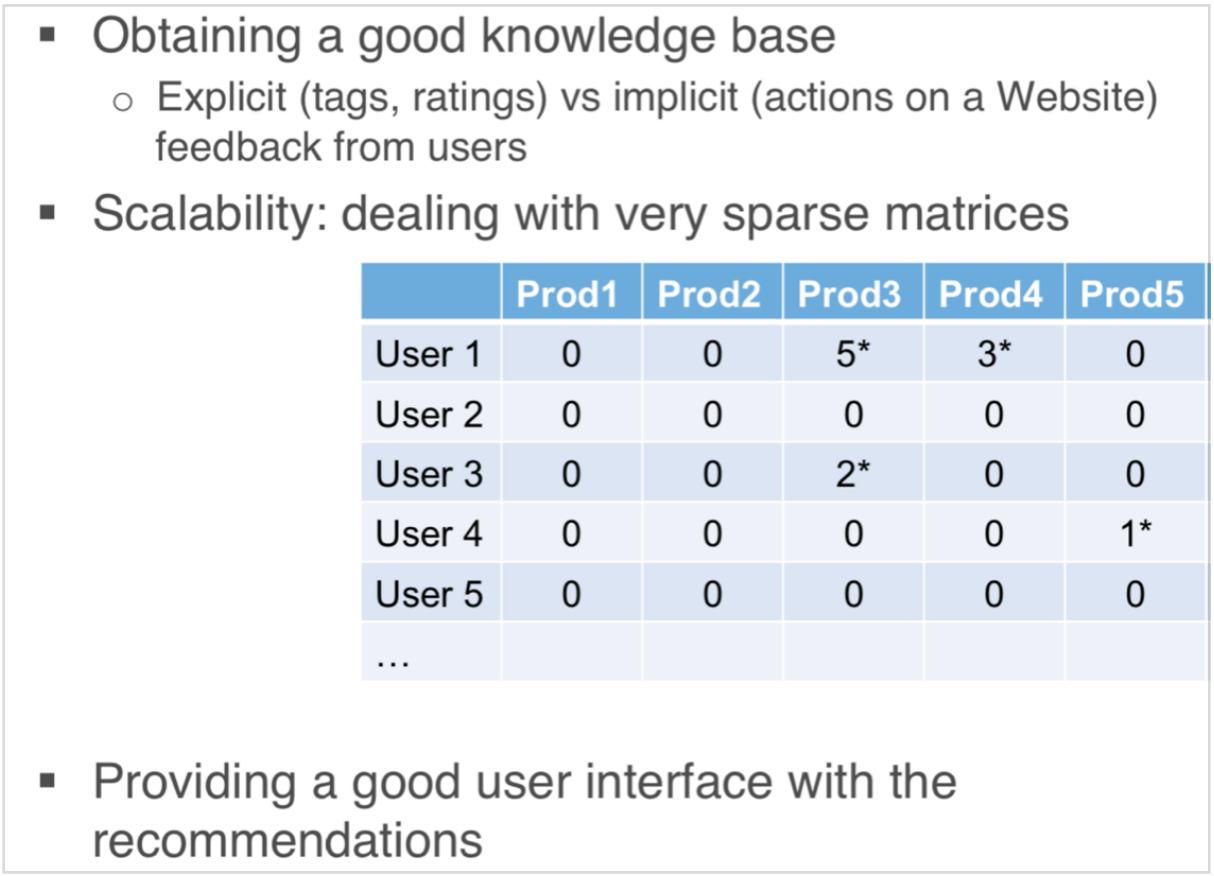
> El más popular , no es una recomendación . Porque no es personalizado para ti.
> Es solo contar

Esferas de colaboración
Tu + lo que has leído.

- Evaluación
- Obtener una buena base de conocimiento
  - no va a ser contenta
  - explícitos
    - tags
    - ratings
  - implícito
    - acciones , feedback de usuarios
- Escalabilidad : Matrices gigantes sparse
  - Web science ➜ Escala
    - Solo tengo 5 items
      - necesitas recomendación
    - Los algoritmos tienen que ser
  - Sociología ➜ Personas
- Buena UI para conseguir contenido
  - ya no es tanto problema, porque la gente está más acostumbrada
- Cold Start
  - Si eres nuevo en un sitio, la recomendación irá a uno por defecto
    - el más popular
    - gente parecida para ti
  - Si el item/hotel es nuevo? , no obtiene reviews
- Gray sheep / Ovejas negras
  - Personas que no siguen el patrón
- Usuarios maliciosos
  - para perjudicar o beneficiar

En las recomendaciones tenemos en cuenta:

- Ventas
- Ratings



	Prod1	Prod2	Prod3	Prod4	Prod5
User 1	0	0	5*	3*	0
User 2	0	0	0	0	0
User 3	0	0	2*	0	0
User 4	0	0	0	0	1*
User 5	0	0	0	0	0
…					

En manchester Óscar creó LikeCube®
CEO era antropólogo , tenía las ideas

LikeCube Limited para toptable ltd.

Innovación en recomendaciones

Presentación de 2007

- Imagina que vas a parís, cómo eliges dónde ir
  - hoteles y restaurantes
  - real travel, hoorah
    - el creado de real travel es el creador de SIRI, es experto en ontología
- Challenge
  - content growth

        - FB añadiendo 100K
        - 10GB of UGC is generado diariamento
    - pero en qué confiar
        - Gusto de los vecinos
- Necesidad
    - confiar al reviewer + 90%
        - confiar en tus amigos
    - dar reviews
    - **encontrar precios y relevante**
- Solución
    - accurate personalizado , filtrado en
- Tecnología
    - Usuarios generado contenido + collaborative filter + semantic web
        - entendiendo
        - taxonomías de gustos
- User journey
    - Filtrar con nuestros test neighbous
- potential add-ons
    - reviews and ratings
- Beneficios
- Ofrecen
    - boost user generated content
    - explotación de amigos, tags

Dos técnicas principales

- Recomendación basadas en contenido
    - generamos perfiles para usuarios y para items
    - Items que has comprado
    - Ejemplo
        - si compras libros, se mira el metadata, incluso el contenido-texto dentro del libro
    - Convertimos personas en vectores
        - embeddings
    - Los items se convierten también en vectores
        - colección de usuarios que han comprado ese item y sus ratings
    - Los hoteles no están usando mucho las recomendaciones
- Filtrado colaborativo
    - No importa las características de las personas ni de los items
    - Estadísticas básicas. Contando

- Cocurrence of items
- Gente que coprodujera , también compró eso
- Tienes que hacerlo directamente, sin librerías

Actualmente usamos herramientas híbridas

a18-04- Recommendation Systems 2

Sesión 29 de noviembre
En teleco , instituto de panel solar
Openscience

- [ ] Próximo día traer ordenador

Gorro investigador para justificar si algo funciona o no.

Debatiremos sobre ¿Cómo hacer una evaluación?

La próxima semana Carlos Badenes (estudiante de doctorado)
Mucho tiempo en la empresa
Modelos probabilísticos de tópicos, evolución de bolsas/vectores de palabras

- [ ] intentar leer los papers recomendados

> Cuando hay mucha subjetividad funciona muy bien
> Por eso en política es un sitio donde funciona muy bien
> Sobre cosas objetivas es más difícil

Dos tipos de sistemas de recomendación

- Collaborative filtering

    - no nos
    - las cosas que estoy recomendando me da igual
    - solo nos centramos en la matriz usuarios / items
    - People who bought this also bought that
- Content-based recommendation
    - cajón de sastre
    - análisis
    - enriquecemos la matriz
    - metemos algo sobre los items
        - analizar el texto de los libros
    - SUELEN SER HÍBRIDOS con collaborative
    - En Content based ➜ Hay arte, metodología , pero no hay ninguna librería
    - Agrega descripciones de esos items

> No hay librerías puras de recomendación
> Hay metodologías
> Hay mucho arte
> Por las características de cada problema que queremos filtrar

Ventajas inconvenientes y características que debemos tomar

En la evaluación, es el contenido interesante

Survey o Review

Hay revistas que se especializan en realizar surveys cada 5 años

> Pasquale Lops et. al., “Content-based Recommender Systems: State pf the Art and Trends”, Recommender Systems Handbook, 2010

Suelen ser los más citados
ACM computing surveys o review

Vision papers

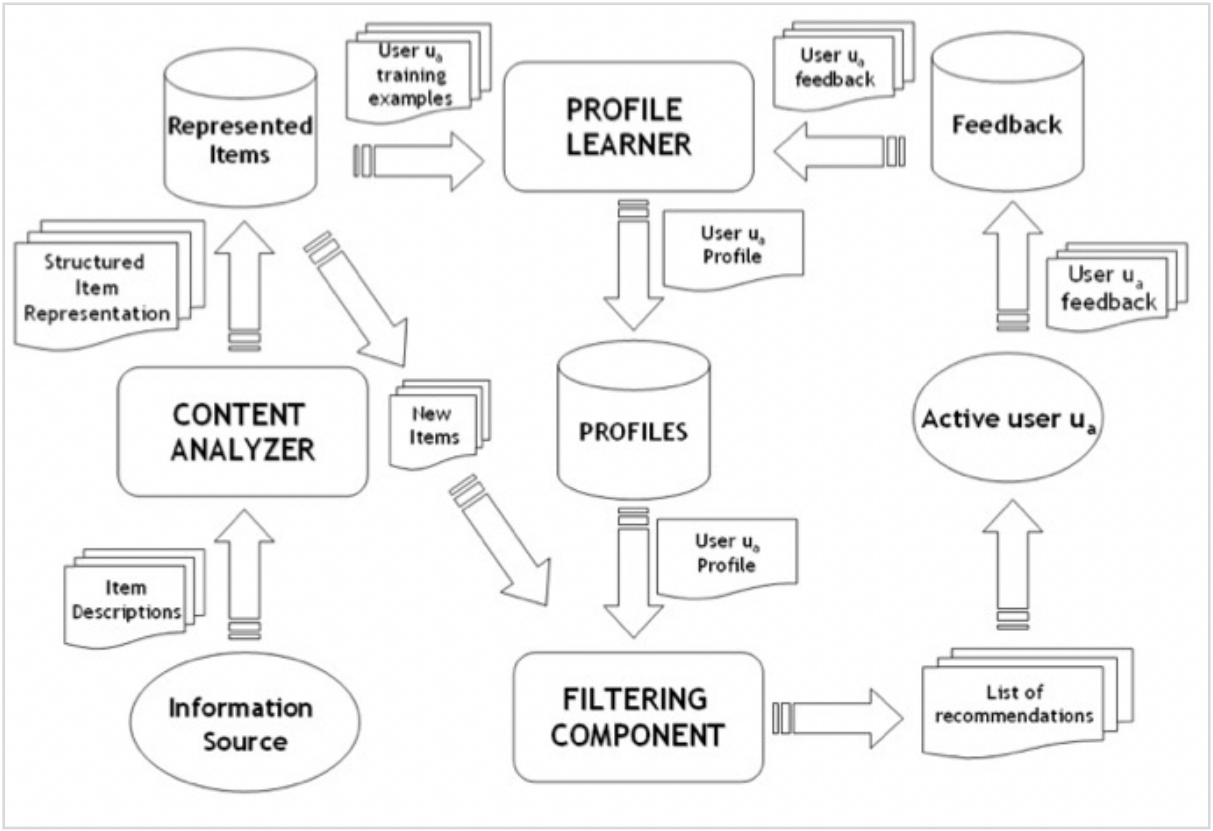
En área de web semántica. Web semantic journal, en enero de 2020 saldrán vision papers , diciendo que en los próximos 5 años debemos ir en este área
Es interesante porque nos dan challenges

> El éxito es combinar un par de surveys + vision papers para tener una visión global, pasado y futuro

Conferencias fundamentales, ahí se discuten

Papers normales, describen experimentos

Arquitectura típica



➜ TF/DF (term frecuency / document frecuency)

- inverse document frecuency
- aparece mucho en un documento, y no en el resto ➜ término muy descriptivo

- Fuente de información
  - IMDB, revistas,
- Content analyzer
  - textos (mejor trabajar a nivel de palabra)
    - sinopsis, reviews ->analizarlo con NLP
      - bag of words ->representamos palabras y documentos, frecuencia de la palabra lematizada
        - se eliminan stop words, director no será
        - verbos , nombres , adjetivos, adverbios (en sentimiento califican)
      - n-gramas (precursor word-embedding) (nivel de letra)
        - letras en el corpus de pla
        - trigazas, combinación

        - son útiles para determinar si dos cadenas son iguales
            - variaciones de palabras: a coruña , la coruña y coruña a
                - las ventanas de trigramas.
        - trabaja a función de letra
        - para corregir palabras
        -
    - **word-embedding**
        - operaciones con vectores

> "Quisiste decir de google no es una recomendación"
> Los correctores de palabras no son recomendadores

- Al analizar el contenido convertimos , items en vectores
    - estructuramos el contenido
    - conseguimos items prespresentados

Profile Learner

Con información sobre usuarios y cómo reaccionan ante nuestro contenido

> Aprendemos perfiles
> Con feedback, conseguimos reforcement

- Podríamos aplicar clustering (descriptivo) y se podría convertir en predictivo

Contamos con y tenemos un

- input: item
- supervisado
- user profiles
    - usuario= vector de items / conceptos
    - si un usuario mira muchas películas de terror
        - en su perfil aparecerán muchos términos de terror: miedo

Si llegan `New Items` no tenemos problemas,
Con collaborative recomendation sí tenemos el problema de cold start
Con content recomendation no, si llega una película de terror, te la puedo recomendar

Filtering Component == RECOMENDER

- Cualquier algoritmo que nos permita detectar similaridad

- vector space models
  - consenos sobre vectores
- naive bayes
- nearest neighbours
- rule-based induction
  - no funciona en todas las áreas
  - árboles de decisiones (que podemos explicar) ➜ XIA
- …

> ¿Es importante dar feedback de por qué?
> Factor psicológico, me gasto mucho dinero en una casa, quiero una explicación

- Tipos de resultados
  - si/no - true/false
    - normalmente los sistemas nos darán un score
    - determinamos el thresehold
  - ranking de recomendaciones
    - es lo que se suele hacer

La evaluación es diferente para cada tipo

Ventajas content based

- It does not take into account other users (the model for every user is independent from the others)
- Recommendations are easier to explain
- Cold-start problem for new items is easily handled

Desventajas content based (Es arte, feature selection,)

- Content analysis may be limited if not enough information available
  - Too dependant on the selected attributes
- Overfitting
  - Recomendaciones obvias, no se valora
- It does not consider the wisdom of the crowd (collective intelligence)
  - Cold start with new users
    - se suele preguntar al principio para categorizarte en un cluster

Introducir outlayer, introducir novedad, no siempre recomendar lo mejor.

Collaborative filtering

Hay dos tipos

- **Memory based**
  - usan/explotan toda la base de datos ➜ LENTO
  - un usuario llega
    - explotar correlaciones entre usuarios
      - damos la vuelta a la matriz y los usuarios son nuestros futures
    - usuarios como features
    - correlaciones
      - pearson, spearman, kendall
  - cogemos los k-most similar
  - agregas información, unes los vectores
    - hacemos una recomendación basada en los k usuarios más parecidos a ti
  - Son recomendaciones de clusters
- **Model based**
  - predecir una celda
    - cada una de estas celdas es un problema a resolver
    - queremos predecir cada una de estas celdas de la matriz usuarios/items
  - lenta para entrenar, rápida para predecir
    - cada día o semana cachea , precomputa todas las celdas
    - no suelen ser sistemas de tiempo real
  - avoide overfitting

Ventajas Collaborative filteriing

no hace falta hacer content analysis para recomendar, da igual que tipo de producto es

Desventajas Collaborative filteriing

- cold start para items y usuarios
- recomendación son más difíciles de explicar (innovative user interfaces)
  - explicar puede implicar los gustos de otros usuarios

Cómo evaluamos los resultados

> las evaluaciones tienen que tener una hipótesis detrás

> definir **baseline** distribución, datos con los que trabajo (filtro usuarios que recomiendan?

cojo todos los usuarios? …)
no podemos tomar todos
no podemos mezclar , los que caen en esa película por accidente + los que caen por tu recomendación

assumption : la clave para que tu evaluación sea válida. Normalmente tenemos solo un disparo
distribución del visitando y el rating explícito sobre películas es similar antes y después
Antes:
con recomendación: 20k vieron la peli, 5k dieron opinión
sin recomendación: 30k sin recomendación , 10k dieron opinión
Después:
con recomendación: 25k vieron la peli, 7.5k dieron opinión
sin recomendación: 30k sin recomendación , 9k dieron opinión
Hay que aplicar
Nos gustaría aplicar un gold standard ➜
Cómo se construye un buen **gold standard**

La evaluación AB-TESTING

Contraste de hipótesis (estadística)
T-Value
si realmente son

¡IMPORTANTE! ESTA ES LA MARCA DE ESTE MASTER
poder explicar nuestras conclusiones con el método científico

Hipótesis
1. el *algoritmo x* SISTEMA (el UI y otros componentes) permite mejorar la percepción de los usuarios sobre los sistemas reocmendados
2. Conseguimos más anotaciones explícitas , se evaluará de otra manera
    1. la evaluación será contar
3. El algoritmo X predice el rating que el usuario Y va a poner el sistema con mayor precisión
    1. binario ➜ precisión, recall
    2. POSIBILIDAD Gold standard

Social, Psicological science ➜ evaluación cualitativa mediante tests de usuarios
Computer science ➜ evaluación cuantitativa

- La construcción de un gold standard, nos permite aplicar nuestros algoritmos todas las veces que quiera.
  - determinar si nuestro sistema predice los ratings de la gente que va a dejar
  - POSIBILIDAD de Gold standard

Evaluación
BBDD
80% training
20% test
K-FOLD Cross validation

Otra opción podríamos considerar el tiempo para hacer la división de
Técnica de embeddings para detectar el orden

¿Cómo crear el gold standard?

Quiero tener el userX - Item Y ➜ predecir el Z (true/false *binario* o % *regresión*)

- GS1 - Me interesa el valor concreto de la predicción del rating
  - comparar la predicción con el resultado final
  - y hacer una agregación del error (sum square error)
  - mi gold standard es la propia base de datos
  - se suele aplicar mucho para el cold star problem
    - un usuario acaba de llegar
    - tiene pocas recomendaciones, tú puedes prerellenar , estimar otra ratings para
  - calculamos el `rock curve`
- GS2 - Me interesa saber que una película le gusta más 1 que a otro. Nos interesa el orden, el ranking. Orden de todas las películas para un usuario dado.
  - nos permite predecir el ranking
  - como google el orden de las "preferencias" / "búsquedas"
  - lo importante es olvidarnos de los números
  - y evaluar lo que podemos ofrecer está en linea con lo que el usuario quiere
  - Crearíamos una nueva base de datos Prima, para tener una
    - para cada usuario, una lista ordenada de preferencias
    - El gold standard será la BBDD y nuestra BBDD prima que contiene los vectores ordenados auténticos

        - Pasamos el problema numérico de square error a ➜ Information retrieval
    - se pone un límite a la lista ordenada
        - 1 ➜ el que más me gusta
        - 5 ➜
        - 10 ➜
            - por google
    - distancia de vectores
        - distancia ordenada, mirar el orden concreto
            - técnica de comparar vectores ordenados ➜ **Jensen-Shannon**
        - podemos comparar el vector como un set
            - Precisión and recall : He devuelto el resultado
            - P@1 he acertado en la primera peli 0,1 ? , P@5, P@10
                - en cuantas he acertado
        - se acerca más a la parte psicológica y será fácil de comprobar
        -

> problema de recall ➜ lo que estamos recuperando es todo lo que podemos recuperar?

Sentiment analysis ➜ veremos *

Evaluación, datasets y challenges

Trends

- Trust: confiar
- recomendaciones basadas en grupos
    - es más difícil de evaluar
    - cómo hacer una recomendación que les gusta a todos
    - 5 libras por booking en un restaurante
- Multi-criteria
    - usuarios - items
    - en vez de matriz , tenemos un “cubo”
    - limpieza, ruido …

- User context: podemos tener más datos
- Explicaciones
  - XIA - normativa para explicar
- Explotación de grafos
  - no tensemos, un grafo completo, item/user context
  - cross modal
    - imágenes + textos ,
- semantic tags/folksonomies
  - y añadir más contexto a la información
- re-rating
- analysis de otros tipos

Conferencias

Conferencias top 1 ➜ **ACM RecSYS**
CHI ➜ Computer Huan interaction

WWW Conference ➜ Knowledge graph

> En web science , los experimentos no suelen ser reproducibles
> Recomendación : empezar y mejorar el gold standard, y luego pasar a las asunciones, y finalmente test cualitativos de usuarios.

a18-05

- carlos badenes
-

How similar are these texts?

- Words + Freq
  - Pros
    - easy
    - confident

    - Cons
        - - semantic
        - +++
- character + Freq
    - Pros
        - multilingual
        - mistake
    - Cons
        - features is constant (número de letras)
- n-gram character + freq
    - Pros
        - easy
    - Cons
        -

Bag of Words

Cada texto es una bolsa de features

- Tenemos que identificar las unidades mínimas del texto `Tokens`
- Decidir trabajar a nivel de palabra o frase
- Tokens
    - su problema son multi-words
        - Inteligencia Artificial
        - My tokens ;) are the best
            - ;) no se entiende como un emoji, se separaría
- Definir, Describir y Eliminar **stop words**. Palabras que no tienen relevancia porque son muy comunes
    - Si no queremos definirlas, podemos usar esta lista: github.com/stopwords-iso
        - hay que revisar la lista para adaptarla a nuestra temática
- **Stemming == Lematizar** : unificar las palabras según su variación de número, genero, tiempo. Cogiendo solo el core lema de las palabras
    - Rules-based
        - podría ser usado para multiidioma
        - es más eficiente
    - Dictionary-based (lemmas)
        - asdf
    - N-Gram Steamming
        - Ejemplo
            - broomstick: *b,br,ro,oo,om,

        - thresholdj
    - El lema sería el número mínimo
    - Lo malo es que podemos obtener lemas que no son palabras válidas

- Binary Bag of Words
    - si está la palabra, pongo un 1
    - Si no tiene la palabra , pongo un 0
    - Hay muchos 0s
    - Lo malo es que perdemos la frecuencia de las palabras, solo nos fijamos en si aparece o no

- Term-Frequency Bag-of-Words
    - en vez de poner 1-0, pondremos el número de frecuencia. Número de veces que aparece una palabra en el texto

- Scaled TF Bag-of-Words
    - usamos una escala, frecuencia/nº de palabras

  Consideramos también las frecuencias del resto de documentos. Para identificar la relevancia en el corpus.
- TF-IDF Bag-of-Words
    - frecuencia/nº de palabras * log (nº de textos / nº de textos en los que aparece)
        -

VECTOR SPACE MODEL

Matriz usando lo que queramos (TF-IDF, binary …)

Distance Metrics

- Euclidean
    - terms frequency no es lo mejor
- manhattan
- chebychev
- minkowski
- Mahalanobis
    - si solo usamos solo la frecuencia no escalada, usaremos esta técnica, y no las otras)

- **triangle inequality**
-

Text similarity

Jaccard Index

Considera la intersección

Cosine Similarity

Considera el ángulo

Es la más utilizada

Text Similarity Test Collab

Collab: https://colab.research.google.com/drive/1_afvnz5RhZRe6jbxdIY-AI912JaEiXvv?authuser=1#scrollTo=rdQ4_dVaE1Dw

Compare Relation in books

Reducción de dimensionalidad

- PCA : Principal Component Analysis
  -
- SVD: Single Value Decomposition
  - a

Latent Semantic Analysis (LSA/LSI) [Deerwester et al, 1990]

- Map documents (and terms

No sabemos la entidad

Probabilistic LSA/LSI [Hofmann, 1999]

- Dividimos

Mezcla de componentes como representación de tópicos
El modelo no es capaz de inferir la distribución de
No es capaz de mirar en documentos en los que no ha sido entrenado

Latent Dirichlet Allocation (LDA) [Blei et. al, 2003]

Es generativo, puede predecir en documentos que no han sido entrenados.
Podemos usar esta aproximación en corpus muy grandes

Número de `topics` fijo

Hyperparameters

La caja significa un bucle

- bucle de documentos
- Topics
- Palabras

Con å y ß manejamos las proporciones

> un documento contiene tópicos, porque las palabras de ese documento pertenecen a ese tópico

Hay que definir

TOPIC MODEL

Valor de å

Si tuviéramos 15 documentos y 10 tópicos
> Define como un tópico aparece
å = 100, todos los tópicos deberían aparecer en todos los documentos
å = 1 ➜ 6 tópicos son posibles en
å = 0.1 ➜ solo 1 tópico o menos es relevante en el documento

- å = 0.01 ➜ solo 1 tópico es relevante en el documento
  - es interesante si necesitamos clasificar, coger labels
  - fallaremos, pero

å = 0 ➜ no cogería ningún tópico

Dirichlet Distribution

Necesitamos mirar la densidad. Otros compradores no nos servirán.
KL, **JS**, He, S2JSD
No podemos usar coseno similaridad

Topic Model Test Collab

- `perplexity` es reducido según converge
- `log likelihood` debe aumentar para
- GridSearch
  - obtiene el mejor lda modelo, que obtiene los mejores valores

> Para seleccionar el número de tópicos para buscar o encontrar,

> Las distancias son basadas en
> Si comparamos documentos sobre 2 tópicos

Topic model
- a word es una distribución de todos los tópicos
- Distance metrics en LDA miden el espacio en , buscamos divergencia
- top words in a root topic creado por hLDA son **comunes**
- uniform topic distribution can be obtained when alpha value is high
- metric that evaluates the semantic cohesion of the topics is **topic coherence**
- **LDA works in non-ecludiean spaces → TRUE**
  - lad space is simplex
- the order of words in a document does not matter
  - en bag-of-words doesn't
- document is

a18-06 Named Entity Recognition

- Oscar Corcho
- 20191025

- 01c_NamedEntityRecognition.pdf

Problemas

No hay gold standard para topic model

NER in an NLP pipeline

1. Tokenization / Sentence Splitting
    1. para cada lenguaje necesitas
    2. porque la morfología es diferente
2. Named-Entity-Recognition (NER)
    1. Identificar las entidades. Extraer
3. ### Part-of-Speech (PoS) NN /VB
4. ### lemmatization
5. ### Stop-Words
6. ### External Semantic Resource
    1. inferir con información extra

> Requiere mucho conocimiento, para ajustar los parámetros
> En ambos campos: estadístico, y gramático

Name entity

> **referent term**or **proper name**.

- **Main types:**
    - Organizations
    - Persons
    - Places
    - Temporal units
        - proyecto Phd. añotator ➜ en documentos legales
            - deadlines
            - fecha
    - Numerical units

- **Biomedicine:**
    - Diseases
    - Proteins
    - Genes

- Substances

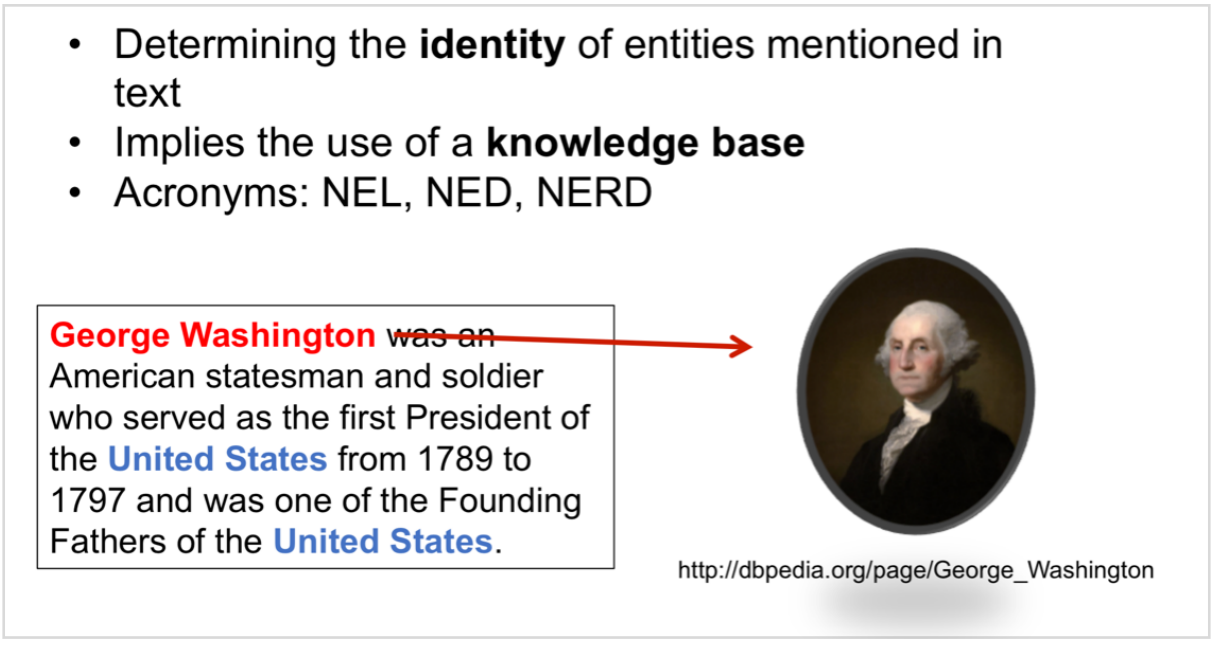
> NER + Clasificación = NERC

1. Aproximación ->Linguistic Models
    1. reglas y heurísticas
        1. MR + {name}
    2. Gazetteers
        1. Diccionarios
    3. Fácil de implementar con librerías y expresiones regulares
        1. usado en panamá papers
2. Probabilistic models
    1. supervised ML
    2. Algorithms

Definitions. Where is NER used?

informationextraction

Named Entity Linking / Named Entity Disambiguation



Co-reference resolution

Cómo nos referimos a una misma entidad con palabras diferentes.

Herramientas para jugar

The one that we developed for ICIJ (Panama Papers)

o https://github.com/oeg-upm/hner-icij

Grammar Rules para detectar las entidades.

Para el consorcio internacional.

Qué papeles hablan de qué personas

Los periodistas, no podían utilizar un servicio online.

A few open services

o Dbpedia Spotlight

- https://www.dbpedia-spotlight.org/demo/ o NLP4Types
  - NER con Linking
  - Natural Language processing + DBPedia
    - Es más que una clasificación
    - Palabras que empiezan por capital case, son potencialmente una única entidad
  - Existe API
- http://nlp4types.linkeddata.es/ o OpenCalais
  - Solo con SVM
  - Identifica la entidad del texto completo
- http://www.opencalais.com/opencalais-demo/
  - Servicio profesional de pago
  -

- Nuria Oliver: Vodafone

La comunidad avanza con challenges.

Un corpus con problemas , y se presentan las soluciones en el siguiente

Conferencia ACL ➜ Asociación de computing linguistics

Describir las evaluaciones.

Los textos se pueden describir como RDF. O como una lista de palabras y entidades.

Sentencias y anotaciones.

Siempre con **Technical report.** para enseñar las evalucaciones

Challenges

Scale, Openness,

- New entities: gente que aparece en twitter. Lugares.
- Entidades desconocidas.

- Los diccionarios son muy grandes para gestionar. Y mucha desambiguación.
- Los textos son muchos. No es un gran problema
  - pero puede ser tratado paralelamente
  - a no ser
    - Prob. Topic model

Pablo 2017 : datos.gob.es

NER - Role Classification Model (RCM)

Linguistic models. Lexical. Heuristics.

> Calleja, P., García-Castro, R., Aguado-de-Cea, L., Gómez-Pérez, A. (2017) Role-based model for Named Entity Recognition. In Proceedings of the 11th International Conference Recent Advances in Natural Language Processing (RANLP)

> SNOMED-CT no es completo
> Proponer nuevos candidatos.

- Entity type
- Entity role
- Role Classification model:

Evaluación

Precisión y Recall

Roles que juegan , las entidades en un texto.

Coger expertos para leer los prospectos y que generan
Reacciones adversas.
Checkeaban con 3 médicos.
No sabes si están haciéndolo en serio o no. Y si el problema es objetivo.

Los expertos van a identificar siempre entidades diferentes.

Anotaciones:

Todos los documentos deben ser anotados al menos por 3 personas.

Es un corpus que puedo utilizar como ¿Gold standard? == ¿Es esta tarea objetiva?. ¿Hay un porcentaje suficiente de acuerdo en las anotaciones?

- `Fleiss-Kappa`
  - [0.5 - 1] ➜ mucho acuerdo
    - los casos raros se tratan
  - [0, 0.5] ➜
    - 0 == Random == Kappa , M
    - no lo uses, no es un gold standard
  - [-1 .. 0] ➜
    - es totalmente subjetivo
    - posiblemente podemos pasar a recomendación personalizada

- Seleccionar al menos 3 anotadores.
- Tienen la tarea de anotar
- Y la colección tiene que ser anotada por al menos
- Mnin task

- Amazon Mechanical Turk
  - control de calidad
- Crowdflower
- Cities At Night

Después del Gold

Panama papers

- Mossack Fonseca
-

a18-07 - Aggregation of preferences: Voting. Social Choice.

Consensus

- Jacinto Gonzalez Pachón
- 20191108
-

Probabilidad | Decisión participativa | Computación Social | preferencias individuales y colectivas.

Unidad 4.
Conexión de ideas y originalidad.
Agregación de preferencias

¿Por qué?

En la web existe una confluencia las dos dimensiones de la ser humano
Dimensión individual y colectiva.
Conflicto filosófica
Ambas dimensiones Muchas veces entran en conflicto.
La web resalta la divergencia.

Genio | Creador ➜
Principio de vasos conectados.

Fascismo, Comunismo

Vivimos sobredimensionamiento de ambas dimensiones.
Emprendedor, Persigue tus sueños, todo a la carta.

Nexo de union grande con estadística:

1. Esperanza (Media) matemática
    1. uniformidad, predice, representa el colectivo
    2. encuentra una parte descriptiva en el todo
2. Varianza
    1. singular, error,

- **Agregación de preferencias**: Construir lo colectivo desde lo individual. A través del consenso.

- Utilizar el Voto,

Autonomía ➜ Variable independiente.
Independencia : Utopía, no es práctica.
Hay un tejido social que lo une todo y afecta a las decisiones colectivas.

La sociedad ➜ Autoridad
El individuo configura la sociedad. Agregando los individuos.

Índice

Voting vs Social Choice

Elementos básicos

- X is la colección finita de alternativas
- m agentes : I = 1..m
-

Regla de votación.

Cada individuo tiene un orden de preferencias.
Resultado de votación: **Elegir el mejor** para el colectivo, para la sociedad.

Nonranked voting system

EJEMPLO
4 candidatos de trabajo

11 decisores

Cada decisor tiene el ranking en su cabeza, pero deben votar solo a 1.
Principio de mayoría en lo mejor. ➜
Si mirásemos la mayoría del peor, ➜

Funeral,
Como se puede combinar lo bueno

Filosofía existencialista.

- Puntos intermedios:
  - Familia
  - Yo me quiero mucho, pero quiero que me quiera otro

Resolver el problema del voto? -> **SOCIAL CHOICE**
Sacrificio va inherente al beneficio.

> Excelencia: etimológico === equilibrio.

Social Choice

Se busca un ranking colectivo
Ranked / preferencial voting system

La agregación de preferencias, resuelven los

EJEMPLO
4 canciones
7 Jueces

> Regla de borda: Gestionar sistema de rankings en una organización compleja.
> Se suman los puntos de todos los jueces, y al mejor se le dan todos los puntos, al segundo mejor.

- Teorema de arrow: No hay ninguna regla de selección social que evite ser manipulada.
  Que evite tener una estrategia detrás de la votación.

No es tan bonito.
Siempre hay un aspecto , una transgresión de

Si descalifico una canción no cumple el **Principio de las alternativas irrelevantes**.

Se podría manipular

UTIL PARA RECOMENDADORES,

Antes lo absoluto estaba en las religiones, luego la política, ahora en el individuo y estamos
Froid decía que los individuos no son tan coherentes. Hay una fragmentación en la que no puede ser absoluto. 2+2 = 4
Quieren saber donde encontrar la verdad. ➜ ESTO ES ABSOLUTO

"Lo ha dicho la máquina" ➜ Esto es absoluto.

La IA debe luchar contra
Detrás hay hipótesis, datos de validación, siempre tienen un bias.

Las matemáticas, el carácter absoluto es un espejismo.
Necesitamos unidades.
1+1 = 2 ➜ solo es verdad en aritmética.

Quien elige las unidades. Tiene le poder.
Una calificación a dos profesores no se pueden comparar.
2 wacas.

1. Unidades
2. Naturaleza / Escalas
    1. Nominal: x != y
    2. Ordinal: x != y, x>y
    3. Intervalar: x != y, x>y , x-y
    4. Razón: x != y, x>y, x-y, x/y

Alternativas irrelevantes: Diferencia entre la naturaleza de los mundos.

Sistemas de votación

- Entre dos candidatos elijo uno
  - mayoría: no hay problema
- Un miembro de varios candidatos
  - normalmente se usa la mayoría
  - Puede haber segundas vueltas
- Alternativas
  - Sistemas de lista
    - D'Hont rule ➜ media más alta
      - ignora la minoría como la moda
    - Greatest reminder
      - leve sesgo hacia la minoría como la media
  - Approval voting
    - Vetar candidatos

Problema clásico

- Colectivo VS Individual
- Principio de la mayoría VS Principio de la minoría
  - Un alumno
    - 10 preguntas de 0 al 10
      - 0000..10 ➜ 10/10 = 1
      - 1111..1 ➜ 10/10 = 1
      - mínimo+mayoría /2 = 0+10/2 = 5
  - Resto mayor

Ley DHont

EJEMPLO Ley DHont (subasta del escaño)

24M votantes , 4 partidos y 5 escaños a repartirse

A ➜ 8700

B ➜ 6800

C ➜ 5.200

D ➜ 3300

Lista A ➜ 1 Escaño

Para el segundo escaño

A ➜ 8700 / 2

B ➜ 6800 / 1

C ➜ 5.200 / 1

D ➜ 3300 / 1

1. Lista B ➜ 1 Escaño
2. Lista C ➜ 1 Escaño
3. Lista A ➜ 1 Escaño
4. Lista B ➜ 1 Escaño

Greatest Reminder

Alternativa: Resto más grande = Greatest Reminder

Leve sesgo hacia la minoría

Approval voting

Mucho peso en lo peor , al contrario del Voting

Se votan todos los que quieras, los que no están marcados son vetados

Hay mucho empate

Social: Situaciones paradójicas. Libertad - Igualdad están en conflicto. Es un dilema.

Social Choice: Agg

Regla de Condorcet Alternativa a la regla de borda

Se cruzan en una matriz. A cuantos elementos domina.

Evita el problema de lo ordinal.

Método de Nanson

Matemático de Alicia en el país de las maravillas

Es el de borda, intentando evitar el problema de alternativas irrelevantes

1. Se suman , se descarta el último.
2. Se vuelve votar
3. Se repite hasta que solo queda uno

Método de Copelan

Cálculo diferencial : Integral (colectivo) VS Derivada(Individual/Partecularización/Un entorno)

C1>C2 … C3>C4

Convertimos la información social para a par.
Si hubiese un bucle, tendría un fallo. Evita bucles, mediante el empate.

C1 ➜ domina a X - es dominado por Y = Z
Se ordena por Z pudiendo empatar.

Método de Cook & Seiford : Compara las distancias.

Establecer unas variables de consenso.
Para determinar los valores de R calculamos la suma de la distancia de cada voto a cada "unidad". Y nos quedamos con la unidad del voto más pequeño.
No se suman. Se estudian las discrepancias por cada voto. Se le asigna el voto con menos discrepancia.

Investigación operativa ➜ método de asignación. Se sustrae el número más pequeño de cada fila.
Con los ceros son las posibles asignaciones.
Puede haber empates.
No te da una única solución, te da una familia de soluciones

> Las inconsistencias hay que estudiarlas (no tacharlas), porque hay información sobre multidimensionalidad. Un dilema.

Teorema de Arrow (1951)

Axiomas

1. Colectivo racional
   - Transitivo
     - evitar bucles
   - Completa
   - no hay duda
2. No trivialidad del carácter técnico
   - dos miembros y 3 alternativas

5. Principio de pareto (a es preferido a b
6. No dictadura
   1. El resultado colectivo no puede coincidir nunca con un individuo o con un grupo redcido de individuos

2. Que todo el mundo sacrifique algo. parte de su individualidad en pro de lo colectivo. Siempre tiene que haber alternativa
3. La verdad
4. Regla de bancarrota: 100K de liquidez
    1. Todos los acreedores siguen siendo acreedores después del

6 Axiomas es incompatible y los sistemas son injustos.

Social Welfare: Carácter ético. Ranking de Intervalo o de razón

> A es 7 veces más preferido que la B

- Una sociedad compuesta por dos grupos etnicos, Yang (mayoritario) Yang. Muy polaricados.

Se lleva a votación el dilema de

- Exterminación de Yang VS No exterminación de los Yang

¿La mayoría tiene el peso suficiente par apedir?

La solución matemática es reducir el conjunto factible.
Acotar con leyes lo que se puede o no puede hacer.
Que eso no se de como opción (política)
El liberalismo no lo hace con leyes.

> **Comparación entre Intensidad de preferencia.**

El sufrimiento de X no se puede compensar ni de lejos el beneficio de Y
El sufrimiento de los yang al ser exterminados, no se puede comparar con el beneficio de los ying.

- **IUC**: Interpersonal Utility Comparisons . Talón de aquiles

Otros temas

> No hay un método justo o injusto. Es el uso que le den a los métodos.

Storytelling - Entroncamiento humanístico

> Para destacar únicamente en hay que entroncar la tradición humanística.
> Storytelling ➜ Narrativa ➜ Tradición humanística
> Una IA Nunca va a poder interpretar los resultados desde un punto de vista humanístico.

Axioma 7 teoría bayesiana. Tema de creencias.
Si AnD > BnD
A|D > B|D
D estando por descubrir, y siendo A más creíble que B
Cuando se sabe D, A sigue siendo más creíble que B

Cuadro de rubens: Aquiles es descubrido por ulises

Gineceo, la madre , travestido. En el núcleo de la confortabilidad.
Aparece Ulises y logra sacarlo de ahí aunque sabe que va a morir.

Contraste de hipótesis, como mucho 0,15.
El sistema pone remedios (noticias) para separar poblaciones antisistema.

> De todo se puede hacer negocio

Decisión participativa : Hasta dónde podemos combinar con métodos computables la unidad individual y colectiva.

> La web va hacia los recomendadores

Próxima semana . Combinación de Dimensiones.

a18-08 - Decisión participativa

- Jacinto
- 20191115
-

Buscar un equilibrio de las dos naturalezas: Lo Individual y Lo Colectivo
Trade off

EJEMPLO
Caso de críticas de cine de Tree of Life.
Película de arte y ensayo. Los críticos dieron muy buena nota, excepto uno.
El público acabó decepcionado.

Tal vez si quieres interpolar

> Principio de la mayoría VS Principio de la minoría

¿Cómo podemos ?

Ejemplo ilustrativo: Alumno con nueve 0's y un 10

> La media aritmética se convierte en una solución de compromiso entre
> el principio de la mayoría y principio de la minoría

Un datasicence suele eliminar las anomalías, la minoría , para conseguir la uniformidad.
[Foto pájaros mirando a un lado excepto uno]
Un caso atípico puede traer disrupción (innovación, mejora, creativo, singularidad)
En lo colectivo está la tendencia.

> Nunca se ha dado un auge de lo individual y lo colectivo al mismo tiempo. Hasta Internet.

Participative Decision Making under satisfying logic

Grupo de investigación: economía y sostenibilidad del medio natural. ecsen.es

Agregar preferencias individuales para definir la preferencia de grupo.

Los individuos, idealizando, son autónomos, e independientes.

Teorema de imposibilidad

Principios sociales + racionales son incompatibles.
Y es lo que queremos resolver.

Seleccionar un método de agregación.

Naturaleza de la información sobre las preferencias. ➜ casuística ➜ Casos

Dando un abanico de modelos.
Vamos a elegir el modelo que equilibre mejor la incompatibilidad.

¿Cómo podemos abordar la incompatibilidad?¿Cómo la vamos a gestionar?

Programación por meta - Goal Programming

> En vez de elegir lo mejor, vamos a conseguir lo satisfaciente.
> Mediante la búsqueda de consenso.

Información sobre preferencias (CASOS)

- Clasificación
  - ordinal - cardinal
    - Ordinal: cualitativo : a es preferido a b. **Social choice** (Ranking completo). ORDEN
    - Cardinal: cuantitativo : a es 7 veces más preferido que b. **Teoría de decisión**. INTENSIDAD de preferencia
      - permite la comparación entre decisores
      - Jing-Jang ejemplo
  - local - global
    - local: método de comparación Pairwise (por pares)
    - global: Ranking, teoría de la utilidad
  - imprecisa - precisa
    - imprecisa:
      - a no es comparable con b
      - a es entre 5 y 7 veces mejor que b
    - precisa
- Representación

8 casos de tratar casos individuales y pasarlos a la colectiva

Matriz de comparación binaria pairwise (por pares)

- Casos 1 y 2 : Ordinal y local
  - precisa
  - imprecisa

  b12 = 1 ➜ a1>a2
  b12 = 0 ➜ a1 ~= a2
  b21 = 1 ➜ a2>a1
  b21 = 0 ➜ a2 ~= a1

  Si ambos valores son 0 , entonces son las a's so similares.

- Casos 3 y 4: ordinal y global

- Casos 5 y 6: cardinal y local: matriz de comparación de valores pairwise (por pares)

- Caso 7 y 8 cardinal y global : pesos de preferencias
  Lo ordinal ➜ binario
  Lo cardinal ➜ escala
  Preciso ➜ número
  Impreciso ➜ intervalo

Buscar solución de compromiso o consenso

Pasar de lo individual a lo colectivo
Global (vectores) - Local (matrices)

> Cook, Seiford - modelo final de distancia

```
\begin{gather*}
U_{1}\left( R^{s}\right) \ =\ \sum ^{n}_{i=1}\sum ^{n}_{i=1} w_{i} |R^{i}_{j} -
R^{S}_{j} |\\
\\
U_{\infty }\left( R^{s}\right) =\min\left\{w_{i} |R^{i}_{j} -R^{S}_{j} |
\right\}
\end{gather*}
```

$$U_1\left(R^s\right) = \sum_{i=1}^{n}\sum_{i=1}^{n} w_i|R_j^i - R_j^S|$$

$$U_\infty\left(R^s\right) = \min\left\{w_i|R_j^i - R_j^S|\right\}$$

Pretendemos minimizar la distancia o maximizar la discrepancia. Coger al representante más cercano a lo social.

Posible entrada electrónica

Donde todo se elige y la única política sería el agregador de información
Libertad + Fraternidad

```
\begin{equation*}
U_{\lambda }\left( R^{s}\right) \ =\ -\ ( 1-\lambda ) U_{\infty }\left( R^{s}
\right) -\lambda U_{1}\left( R^{s}\right)
\end{equation*}
```

$$U_\lambda\left(R^s\right) = -(1-\lambda)U_\infty\left(R^s\right) - \lambda U_1\left(R^s\right)$$

Procedimiento computacional (Operativa)

Hacer lo compatible lo incompatible.

- Cambio de variable
- Modelo extendido de programación por meta (Extended GP model)

- se pretende rebajar la posición del individuo

```
\begin{gather*}
R^{s}_{a} +n_{1} -p_{1} =4\\
n\ y\ p\ son\ variables\ de\ desviación\\
R^{s}_{a} +n_{2} -p_{2} \ =\ 2\\
Sacrificio\ de\ un\ individuo\ en\ favor\ del\ consenso\\
M_{m} \ \psi ( p_{i} ,n_{i})\\
El\ GP\ pretende\ que\ el\ sacrificio\ sea\ el\ mínimo
\end{gather*}
```

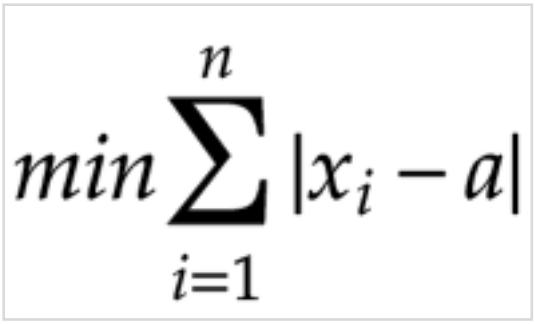
$$R_a^s + n_1 - p_1 = 4$$
$$n\ y\ p\ son\ variables\ de\ desviación$$
$$R_a^s + n_2 - p_2 = 2$$
$$Sacrificio\ de\ un\ individuo\ en\ favor\ del\ consenso$$
$$M_m\ \psi(p_i, n_i)$$
$$El\ GP\ pretende\ que\ el\ sacrificio\ sea\ el\ mínimo$$

Extended GP model - Casos resueltos

1999 - agregar ranking y precisos.

Análisis del alumno que protestó sobre la media aritmética

Mediana está basado en la mayoría.
Valor que minimiza la sitancia `x_i-a`



```
\begin{equation*}
min\sum ^{n}_{i=1} |x_{i} -a|
\end{equation*}
```

Moda: lo que más se repite

MODA ➜ MEDIANA ➜ MEDIA

Media: minimizar la ditancia `(x_i-a)^2`

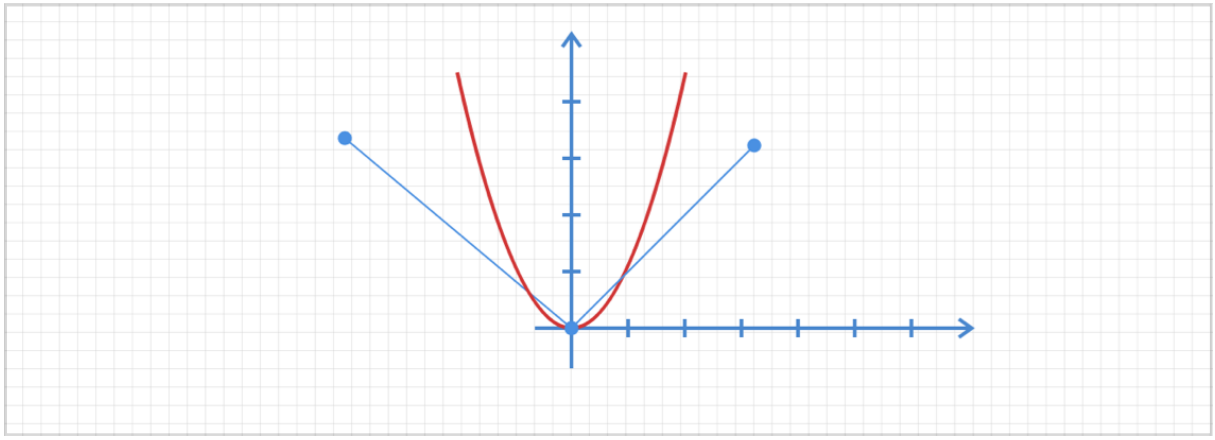
Si se ha abandonado la mediana en favor de la moda.

Medida de dispersión, es no diferencial, es decir, que en un entorno de 0 no podemos reducirlo para convertirlo en problema lineal. (Tengo infinitas tangentes)

La moda si es diferencial, si tiene una tangente.
Al estar elevado al cuadrado, el valor atípico gana importancia.

```
\begin{equation*}
min\sum ^{n}_{i=1}( x_{i} -a)^{2}
\end{equation*}
```

$$min \sum_{i=1}^{n} (x_i - a)^2$$



No existe el agregador perfecto.
Los recomendadores solo utilizan un método de agregación. No contrasta entre

Libertad, Igualdad y Fraternidad.

Prospecto médico: indicaciones y contraindicaciones
Pensamiento crítico: contemplar la duda razonable.
Probabilidad: medida del error y duda razonable (judicial: inocente + pruebas)

IA ➜ riesgo ético.
Nostalgia de absoluto y verdad
Pensar que desde un algoritmo se puede alcanzar la objetividad es un error.
Detrás de ese algoritmo siempre tiene una hipótesis

Paper de Jacinto: The design of socially optimal decisions in

Mayoría = Libertad
Minoría = Igualdad

Solución de compromiso:
La media suele estar bien, pero podemos ajustar y reforzar un lado (mayoría o minoría)

moviendo los lambdas

La docencia es en espiral

Probabilidad: Error + actualización de creencias.
Engañar a la gente es más fácil que sacarlo del error

- [ ] Entregar Lecture Notes: Indicar que hemos entendido en clase

a18-09 03 Social Computation / Computación Social

- Javier Bajo - jbajo
- 20191122
-

Índice general

Índice

Introducción

Overview y current trenes
focus en social computing y sistemas multiagentes

Esquema de la presentación:
Social computing ➜ agent-based artificial societies ➜ virtual organizations ➜ trust and reputation

Objetivos

- Overview de social computing y web science
- overview de maquinas para el modelado de social machines: current trends en sistemas multi agente
- oportunidad de investigaciones y areas de aplicación actuales

Web science and social computing

- la web es la construcción más grande de información humana
  - la web transforma la sociedad
  - la web es
- Web science combina multiples disciplinas como sociología, economía , matemáticas, computer science
- Nos centraremos en **social aspects** y **social computing**

Social Behavior

- Observar el comportamiento social
  - puede ser organizado o auto-organizado
- Si miramos individualmente Comportamiento social
  - Por ejemplo: comportamiento de abejas [Vídeo: Nature Id tags on bees]
    -
    - Es complicado ver comportamiento únicamente individual en este tipo de sociedades
  - En el comportamiento humano,
- Global social behaviours
  - podemos usar algoritmos basados en estos comportamientos biológicos
    - para resolver problemas
    - también utilizados para efectos sociales en películas. World War Z ➜ zombies crean torres como las hormigas.

- Sociedades humanas son avanzadas y tenemos
  - Collaborative engineering
  - comunicación directa, negociaciones , economía internacional, etc
  - emergent complex global behaviours

    - es regulada con leyes, normas, institudciones
    - el comportamiento humano individual está basado en racionalidad, aspectos sociales …

- MABS (Multi-agent-based simulation
    - inspirados en entidades biológicas para resolver problemas humanos
        - Ant behavior to solve urban problems (urban biology)
            - Urban biology
            - Ant behaviour to solve urban problems - YouTube
    - observar el comportamiento humano y simularlo
        - Evitar colisiones, evitar Bottlenecks: A statical similarity measure of aggregate crowd dynamics.
            - Con la simulación podemos decidir dónde poner los objetos etc.
            - Cambiar la distribución de objetos
            - Simular una evacuación
                - realmente es difícil capturar todas las complejidades del comportamiento humano.
            - Videos
                - A Statistical Similarity Measure for Aggregate Crowd Dynamics
                - A Statistical Similarity Measure for Aggregate Crowd Dynamics - YouTube
        - Negotiation in multi-agent systems
        - Virtual institutions for water rights negotiation
            - para capturar

- MAS-based approach (ANTS)
    - swarm intelligence
    - emergence
- Social Behavior - Organizational Theory
- Symbolic Knowledge Representation (HUMANS)
    - Self organizing
    - Re-Organazing

Una sociedad es un sistema complejo, intentamos promover la solidaridad y la estabilidad.

1. Estructura social: Un grafo representando las relaciones entre las diferentes entidades del grupo
    1. Énfasis en la idea que la sociedad es agrupado en grupos relacionados

estructuralmente.

2. Funciones sociales: en términos de
    1. normas
    2. customs??
    3. tradiciones
    4. Instituciones: para controlar educación , seguridad,
        1. sociedades reguladas estrictamente

Funciones sociales : Convenciones que tenemos en nuestra sociedad

Tenemos objetivos que queremos alcanzar.

- En economía queremos definir estados sociales donde la gente intente ser lo mejor posible
- En educación : funciones sociales
    - estudiar en la universidad es “bueno”
    - es “bueno” ser mejores notas

Las funciones sociales pueden ser óptimas para el conjunto social, pero puede ser malo para la minoría.
Ejemplo : enviar personas a prisión, reducimos la interacción con la sociedad, ya no puede contagiar su comportamiento malo. Sin embargo estos individuos no mejoran en la cárcel, sino que empeoran. Al final no pueden ser reinsertados en la sociedad.

Herramientas :

- Mason
- NetLogo
    - bibliotecas con modelos para simular:
        - altruismo, rumores en redes sociales, clusters de votos,

- La clave en sociedades humanas
    - Knowledge representation
    - razonamiento

Psicología cognitiva: teorías basadas en procesos mentales las cuales están basadas en comportamiento racional

- Representación de estímulos externos
- manipulación de representación
- …

- Basic cognitive processes en IA
  - Sensación
    - cámaras, micrófonos, presión , narices electrónicas
  - Percepción: nos permite interpretar el estímulo
    -
  - Atención
  - Memoria

Procesos cognitivos para : Pensar, Lenguaje e Inteligencia

Mecanismos en nuestra sociedad

- Comportamientos sociales
  - interacción
    - dialogue: nlp
  - cooperation
    - working together (teams) to achieve a **common goal**
  - delegation
    - dividir un problema en problemas más pequeños
    - task assignments
  - Coordination
    - cómo resolver problemas indeseables
    - manejo del proceso de resolver el problema
    - resolución de conflictos
  - Negotiation
    - acuerdos que son aceptables para todas las partes

> el reto es reorganizar estos mecanismos para una sociedad artificial / híbrida donde conviven humanos y máquinas

Social Computing

Analizamos comportamiento social
Analizamos sistemas computacionales
En la interacción tenemos el Social Computing

> Usar ordenadores para propósitos sociales

> Web es una nueva herramienta para networking, compartir intereses, publicar pensamientos …

Diferentes Definiciones

- Tom Erickson IBM: sistema técnico que soporta comportamiento social entre las personas y utilizamos este comportamiento para distintos propósitos. Son los mismos comportamientos que
  - Ejemplos
    - Amazon, fue el primero en incluir la evaluación de los compradores
      - Grupos de personas que están computando algo socialmente
    - Wikipedia: La contribución produce una enciclopedia gigante
    - Ebay: Subastas online (Auctions)
      - replica un comportamiento social sobre tecnología
  - Debemos observar el comportamiento

- Leo Von Ahn - Captcha : La tecnología que soporta cualquier tipo de comportamiento social atravesado de sistemas computacionales. (blogs, email , wiki, social networks)
  - Facebook: comunidades online de amigos
  - LinkedIn: comunidades online para buscar trabajos
  - Darpa Network Challenge: la solución viene
    - DARPA and the Red Weather Balloon Challenge - YouTube
    - Intentar definir redes más complejas
    - soltaron 10 globos aerostáticos en estados unidos, cómo podemos localizarlos rápidamente
      - MIT lo consiguió en 9 horas a través de las redes sociales
  - Twitter
    - Estimating County Health Statistics with Twitter - YouTube
      - combinar términos a regiones. Diabetes
  - Flickr: predecir niveles de nieve desde las imágenes de flickr
  - Captcha (Completely Automated Public Turing Test to tell Computers and Humans Apart): programa que detecta a un humano. Test de turing completamente público.
    - leer texto o seleccionando imágenes
    - ayudamos a mejorar palabras escaneadas OCR , para libros o direcciones de calles google maps
  - OpenStreetMap: ayudando a mejorar los mapas
  - Waze: reportando la velocidad
  - JustEat
  - Wayook.es : limpieza de hogar
  - Crowdfunding: Kickstarter | Indiegogo
  - Nicho todavía sin resolver Según Javier: Muerte
    - ataúdes, entierros, flores, cremaciones, …

- Lee and Fischetti : la vida real debe estar llenaa de social constraint - the los procesos de la sociedad de la cual nace. Ordenadores puede ayudar si nosotros usamos a crear maquinas sociales abstractas en la web.
    - Procesos en los cuales la gente hace el trabajo creativo y la máquina hace la administración

  Trabajo creativo es difícil de definir.

- Lo que actualmente hay en la web es:
    - social media,

  Pensar en soluciones para :
- solución de problemas sociales distribuidos
- .
- .

  (La mejor definición)
- **David Robertson** : El poder del ordenador social reside en la combinación programable de contribuciones desde los humanos y ordenadores.
    - los humanos aportan: conocimientos, competencias, habilidades
    - ICT

  En el futuro:
- Dos ejes:
    - Y = Más cómputo
    - X = Más gente
    - Big Data, Big Compute, Conventional computation
    - Social Machines, Social Networking

  Nos movemos muy cercano a los ejes. El reto está en combinar la potencia de ambos ejes.
- Social machine healthcare and disease
- Social oganisation of transport
- Social response to emergencies and crime

- Air traffic control,

- Social computing:

- Aplicaciones:
- Infraestructura tecnológica:
  - web, database, multimedia, wireless , agente,
- Theoretical underpinnings

Areas de aplicación:

- Entidades inteligentes y entretenimiento
  - Huggable Robot Befriends Girl in Hospital - YouTube
  - Alexa: entendiendo NLP y proveyendo soluciones
  - Geminoid HI-4, el robot humanoide más avanzado del mundo - YouTube
    - Hiroshi Ishiguro : Supervisor de Javier en el año PostDoc

Issues abiertos de investigación

- Representación de información social y conocimiento
- Agent-based social modeling
  - mejorar los modelos
- Human-agent interaction
- Analysis and prediction
- Negociación y sistemas de decision making
  - agregación de votos, para generalizar

- Trabajos potenciales:
  - Crear una presentación sobre websicence y social computing
    - human-agent interaction
  - analizar centros de investigación actual y carreras sobre social computing: teoría, aplicaciones, tolos?
  - analizar conferencias: CFPs, papeles aceptados
  - analizar empresas que están investigando o invirtiendo en estas soluciones
    - MIT han recibido 10Millones $ de twitter para centrarse en

- [ ] trabajo en grupo (2, 3 personas) de 10 páginas sobre algo de nuestro interés + presentación de 10 minutos. Primera o Segunda semana de enero. 10 o 17 de Enero. Si presentamos el 10 , el 17 será libre.