AN ESSAY ON

KNOWLEDGE OF IGNORANCE

MARTIN VÍTA

UČO: 333617

IV123 – FUTURE CHALLENGES OF INFORMATICS

FACULTY OF INFORMATICS

MASARYK UNIVERSITY

BRNO

SPRING 2013

Contents

Abstract

Every question or keyword we put into the search field of Google and every question we write in a discussion forum reveal something we would like to know. So, that implies it shows a very small part of our nescience, or rather ignorance. The manner of our web usage also provides interesting data about our knowledge and our nescience. The digital track contains much information which can be helpful when we want to reconstruct someone's ignorance – "complement of someone's knowledge".

In this essay I would like to point out some important aspects of gathering and reasoning about someone's ignorance, suggest some related ethical issues to solve and show how these findings can be useful when building a "smarter world" for everyone.

Introduction

Today's society is often called *knowledge-based society* or *society of knowledge*. Knowledge society is typically understood to be an opposite of the industrial society. Several years ago Konrad P. Liessmann pointed out that this difference is rather imaginary: in fact we deal with information or knowledge in the industrial manner today. Regardless of these approaches we can say that knowledge plays a very important role in many aspects of our life and it is a keystone of our civilization. Despite this the notion *knowledge society* became nearly a buzzword.

Knowledge management became popular in the corporate environment in the last twenty years. Global companies have usually implemented very sophisticated tools for gathering, storage and retrieval of desired pieces of knowledge. These tools allow anyone inside the company to access the (explicit) knowledge of members of different teams in the company and share own knowledge with others – for example, enterprise wiki-based systems are widely used for this purpose. Other software tools inspired by social networks can help employees to find someone who has some necessary knowledge for solving certain problem. Companies also quickly set up relevant strategies, rules and processes how to deal with employee's knowledge. The main objective of such practices is (usually) to improve the performance of the company, provide a competitive advantage and facilitate the innovation process.

Knowledge which was formerly distributed among many members of an institution is suddenly available to anyone *en bloc*. Thus, individuals can flexibly use the whole knowledge of the team. If we incorporate the the terminology of multiagent systems and epistemic logics, we can roughly say that distributed knowledge is "being converted" into a knowledge of concrete group of agents.

If we thing about these principles and tools, we will probably notice one interesting – and not obvious – fact: these systems are used for dealing with "positive knowledge", for example:

> *Employee A knows that employee B has some experience with Oracle databases. Everyone in our department knows that customers buy usually books about PHP programming and Apache server administration together etc.*

In the terms of epistemic logic we can say that negation (if used) does not stand before modal operators of knowledge (only inside of "prepositional" subformulae). It is natural – it is a first

step in dealing with knowledge. A great development in many parts of computer science and related disciplines (e.g. text mining, natural language processing) enables a big opportunity (and adventure): ***reasoning about knowledge of ignorance,*** e. g. agent *i* knows, that agent *j* does not know that *p*. Consequences of this approach are sometimes hard to imagine. In this work we are going to present some of them.

But there is just one widely known example of this phenomena: the legal principle *ignorantia juris non excusat*, literally *ignorance of the law is no excuse*. What can be done when we know, i. e. we have the knowledge that certain group of people does not know some concrete acts? Is there a chance to use this for our business? How? What are the risks? Are there any ethical issues? These and other questions we are going to state and discuss for different domains.

Preliminaries

Before we start we are going to recall some basic notions concerning our topic. Definitions are informal but sufficient for our purposes. Sometimes a brief description of the notions instead of the definitions is provided.

Explicit knowledge is knowledge that has been articulated, codified, and stored in a certain media. It can be (easily) transmitted to others. (This definition was taken from *Wikipedia*, one of the biggest storages of explicit knowledge – according to presented definition of explicit knowledge). Examples of explicit knowledge can be found in manuals, encyclopedias etc.

Tacit knowledge is often understood as opposed to explicit knowledge. It is knowledge that is difficult or rather impossible to articulate it, formalize it and transfer it to another person. The concept of tacit knowledge was introduced by Michael Polanyi in 1958 – the main idea of this concept is based upon the assertion "we can know more than we can tell." CITE. *How to* search some information on the internet is a good example of tacit knowledge.

It is possible to say that tacit knowledge can be characterized as "know-how", while explicit knowledge by "know-what". There is also a "grey zone", sometimes called *implicit knowledge* – it is a knowledge which has not been formalized (yet) but it is *possible* to formalize it.

Multiagent epistemic logic is the apparatus which allows us to describe the knowledge of an agent or of a group of agents in terms of propositional modal logic. We will use a logic which contains four modal operators with the following meaning:

- $K_i p$ – Agent *i* knows *p*.
- $E_G p$ – All of agents of group *G* know *p*.
- $C_G p$ – All agents of group *G* know *p* and know that they know *p* etc. (common knowledge operator)
- $D_G p$ – *p* is a distributed knowledge for agents in group *G*.

Comprehensive overview of multiagent epistemic logics is provided in [1].

Basic Principles and Notions of Our Vision

As mentioned in the Abstract, most of keyword searches can be interpreted as *ignorance* in some context: if someone types “second largest city Southern Moravia“ into the Google search field, we can infer that the user probably does not know what is the second largest city of Southern Moravia. If someone (Czech) use a web dictionary for translating the noun “city“ to Czech, we can infer at least that he or she does not know the meaning of the word “city“ in English, but most of us probably add something like “his or her English is poor“.

Analysis of our questions or queries based on a simple assumption (we ask in case when we don§t know something) can provide us extremely important data. Of course, the reason of typing “city“ into the web dictionary can be different – we can imagine a linguist who controls the quality of the dictionary, but this seems to be a rare event.

This kind of ignorance we are going to denote by the term *explicit ignorance* (for the purpose of this text only! In other text this notion can be used in different meanings.) – the sentence about the ignorance can be transformed by simple syntactic rules from the set of keywords: second largest city Southern Moravia → Agent *i* does not know the name of the second largest city of Southern Moravia. (→ $\sim K_i$“Znojmo is the second largest city of the Southern Moravia“)

It can be seen that handling of explicit ignorance may be relatively simple (if we have an NLP toolset providing relevant functionality). Any form of question is directly linked with an expression about ignorance.

A different situation is when analyzing user's ignorance from posts in discussions bellow blog articles, Twitter contributions etc. Let's illustrate the idea on a simple example. Some person put his or her posting under the article about taxes. He or she says that governmental revenue is getting bigger if the taxation rate increases. From this posting we can conclude that he or she doesn't have a basic economic background. More preciously, he/she has no knowledge about Laffer curve.

This kind of ignorance we are going to denote as implicit ignorance. The evidence of the ignorance can be deduced (even by machine) from the content (of a collection) of sentences and the context. It is possible also to establish the notion of *tacit ignorance* – but without purposes for this text.

Before the internet era, ignorance of a given person was distributed among many people: teacher of maths knew that particular person had no knowledge of analytical geometry, accountant knew his customer had no knowledge about taxes, a doctor had evidences about someone's knowledge and ignorance of medicine etc. For this age it was typical, that “holders“ of knowledge of someone's certain ignorance didn't communicate with each other. Sharing of these information was impossible and collecting pieces of knowledge of someone's ignorance was done only in rare cases and with relatively big costs. It was rather detective or investigative work, closely tight with particular persons – no mass action.

In the internet era these things have rapidly changed. Among any aspects I would like to point out following four:

- Exploratory analysis of big data becomes relatively cheap for many institutions and individuals.

- Today, a big amount of personal expressions is available to worldwide public (e.g. posts on social networks, discussion forums. Email communication is available for internet providers etc.).
- Recent advances in textmining, web mining and natural language processing allow us to gather more valuable information from the data then in the past.
- Rising impact of personalization.

If somebody in the pre-internet era tried to get ignorance of a person in question, he had to find out people with some relation to this person and had to ask them directly (or indirectly). Now it is sufficient to follow the "digital track" of the person in question. The digital track provides us many evidences of someone's implicit ignorance.

It is important – but not obvious from the first point of view – that "missing values" (i. e. certain undiscovered ignorance in this case) concerning the person in question couldn't be reasonably estimated in the pre-internet age. Now in the age of big data we are able to guess these missing values from thousands and thousands of analogous instances.

Since now we are able to deal with people's ignorance "in an industrial way" – globally and massively. We are able to generalize our knowledge about some group's ignorance and conversely, we can use these general discoveries in personalization of different (web)services. We are going to illustrate these principles in one of the next sections.

Key ingredients: data mining, text mining, web mining and natural language processing

From a certain point of view the task is that we have a large collection of data – mainly texts on the internet linked with large amount of persons – and we would like to discover them with respect to find out their personal ignorance. Hence we have two subtasks:

- Classify the persons whether they know or don't know some fact (theory etc.).
- Find some frequent patterns in data about ignorance, for example: "if some doesn't know the meaning of words like "Sunday" and "Monday", he won't probably know how to create a future tense of "be".

The role of natural language processing is clear. In one hand we must use some low-level issues (like segmentation, named entities recognition etc.), in the other hand we must use high-level or complex topics (like tectogramatic tagging etc.).

Rule of Association Rule Mining and Machine Learning

Widely used method for discovering interesting relations between values of attributes in large databases, typically in the form of implications. Today, there is a plenty of algorithms implementing this functionality. This functionality became very popular mainly thanks to the usage in so called market-basket analysis: the task is to find which goods are likely bought together, for example: if someone buys meat and potatoes, then he or she buys also a beer at the same time.

Association rule mining is analogously applicable on our tasks: we would like to produce implications like “if someone can't understand complicated formulations and can't calculate with percents, then he or she is not able to differentiate disadvantageous loan offers from the good ones.“

Supervised methods of machine learning are – and probably will be more and more widely – used for classifying web users into groups – in our case by several criteria concerning ignorance and knowledge: “does this person belong to people with analytical thinking and lack of knowledge of history?“ etc. These classifications allows us to provide more “fuel“ for mining association rules.

Obviously, these technologies and data from the internet can be both used for improving our knowledge and lives and also abused. In the following section we are going to demonstrate the power of both sides.

Four Examples of Dark Future

Case One – License Agreements and Other Adaptive Texts

In many cases when you buy something, the contract between you and the provider is formalized by some kind of “terms and conditions“ or “license agreement“ etc. These documents usually specify the extent of the fulfillment, price conditions and other important clauses. Today, end-users typically should agree with the same terms and conditions as an arbitrary similar client. If you want to set up a new account at a new bank, you will probably get the same list of terms and conditions as your neighbor. There is no personalization behind this process.

In the new-internet era this will be changed. These texts can be on-the-fly adapted to your current level of knowledge and ignorance, hence there is a big space for “companies“ which have their business based on “soft-cheating“. Common today's principles of obfuscation, distortion and language confusion will be improved on an unimaginable level and used according to current state of customer's ignorance.

What can you do with client who can't understand slightly difficult text and can't calculate with percents? If you are a honest merchant? And if you are a unscrupulous dealer of non-bank financial institution offering loans? What could you do in the situation when you can address these people in thousands?

In future, due web mining capabilities overcome the common trade-off between personalization and costs. World of blanket cheating – but highly personalized.

Case Two – Healthcare

Public health issues belong to one of the most important topics in election campaigns, pharmaceutical industry is more and more powerful and influential. Although there are experts in this field (i. e., mainly doctors), many of us often compare the expert's diagnoses with experiences of other people on the internet, many of us are trying to find the diagnosis

without seeing the doctor's. Google already experimented with predictions of disease spreading via analysis of sequences of queries linked with spatial data.[1]

It is not so difficult task to decide whether the user is has education in medicine or related fields or at least some biological background – if you have enough examples of his or her queries, contents of mails and lists of visited websites. From posts in discussion forums or on social networks can be predictable whether the person is predisposed to believe prejudices or whether he or she looks with favor upon pseudoscience.

Again, similarly as in the first case, e-drugstores or e-pharmacies can use this knowledge about your ignorance of certain alternative medicaments or alternative methods of treatment to recommend you more expensive and long-lasting way of therapy.

This topic was slightly inpired by [2].

Case Three – Elections and Politics

People in political marketing have been using various results of webmining probably since the beginning of this discipline. It's mainly a case of application methods of sentiment analysis on this domain. Political parties want to see the impact of a certain step on the public opinion or want to estimate the popularity of a particular member of the party etc.

Deep analysis of ignorance of wide mass of electorate can provide an unprecedented opportunity to "complete the knowledge of certain groups people in a desirable way". More formally, the task is to provide voters a maximal set of opinions which is consistent with the current knowledge of the electorate and also which with the party's program. If someone find out that certain voter doesn't know particular economic law, then the party can serve him or her some simple populist solution (although it can be predicted it won't work well). If someone doesn't know the causes of a certain affair, than we can provide him or her our interpretation (according to our goals) of this affair.

Case Four – E-shopping

Business was one of the most important initiators of the applied research in datamining and later, webminig. Sophisticated methods of market-basket analysis were developed and are widely used among internet merchants. Probably every internet user encountered results of recommendation systems. Behind many of them there is a non-trivial application of machine learning methods, data analysis etc. Despite of this, there is a gap "on the client-side". Users usually see the same product detail page (in except of some features – e. g. recommendation of other goods) as the other visitors of considered e-shop.

Let us imagine an e-shop with electronics and a user looking for some kind of device. He or she is just on the way to the page containing details about it. If we employ the analysis of user's ignorance in the process, we can personalize product detail pages of the e-shop in a more complex way – if the systems knows that the user doesn't know that goods of other manufacturers are cheaper and have better parameters, then we can generate the page full of announcements like "the best choice in this class" etc.

1 It is possible for anyone to experiment with Google prediction API via: https://developers.google.com/prediction/

Four Examples of Bright Future

The sketched visions seem rather pessimistic. Let us note that this is the typical way how mainstream media describe the usage of advanced technologies. But is necessary to realize that probably each of the cases above has its "positive" counterpart. Let us introduce revised cases from this positive point of view.

But at first, we are going to start in a slightly more general setting.

Every ignorance can be regarded as an opportunity to an action: for "the holder of the ignorance" it is the opportunity to learn, for "the finder" it is a chance to correct the state – to provide some inputs for educational process (or typically provide the relevant knowledge).

Case One – Adaptive License Agreements and Manuals

As mentioned above, license agreements, terms and conditions can be on the one hand understandable and clear and on the other hand incomprehensible and unclear – this is partly the property of the text, but partly it also depends on the user. Today, most providers of several services or producers of different goods use some universal versions of terms and conditions for each customer. Thus, for one group of users are the terms and conditions so complicated that they do not understand them, for others they are needlessly too wordy.

Analogous situation we can find among different manuals to anything: the quality and usefulness of the manual is not something absolute – it strongly depends on the target group for which is the manual intended. The same manual can be treated as bad because it is too brief (for beginners) and it can be treated as very convenient for expert because it doesn't contain trivial advices as well.

If we know the ignorance and knowledge of a particular user, we can prepare a manual, which perfectly fits his or her needs. In many cases there is no reason to have one manual for all. The style of writing is also interesting: if someone prefer short, clear sentences, we can satisfy him or her of course. If someone doesn't have problems with long, complicated compound sentences, we can provide him or her this form of the text.

This improvement leads towards extraordinary time-savings. Advanced users are not forced to read things they already know, beginners avoid long-lasting search of unclear terms and statements. It is obvious that this savings can be recalculated by means of money.

Aspect of motivation is also relatively important. Thanks to adaptive content of the manual, people will realize that study of manuals is not wasting of time. If someone is forced to use manual which doesn't correspond with the level of knowledge and ignorance, his or her motivation to use the manual quickly falls to zero.

Case Two – Effective Communication in Healthcare

Data about people's ignorance in medical issues can be very useful when modeling the spread of some disease. Informed people with relevant medical knowledge behave in a more responsible manner and do not "add more worries to medical staff". This is – in some sense –

global, public usage of these data.

From the point of view of a particular internet user the processing of knowledge about "medical" ignorance and its positive application can avoid situation described like "If I knew that ..., I would have done something else...", for example, ask for a special laboratory examination, use another preparation etc. Analogously as in the previous case, package instruction leaflets (of course in an e-form) can have content adapted to user's current level of knowledge of medicine. Used language and used notions can be chosen accordingly to the user's communication manner. Therefore these leaflets can play the original role – to be a convenient advice for the patient.

The other opportunity is to integrate the "personal ignorance processing" functionality to doctor's expert systems. The application can suggest the doctor to explain something what the patient probably doesn't know.

The last situation we are going to mention is ranking of results in a web search. After the application of this strong kind of personalization the final order of results can take in account the level of knowledge of a particular user and also it can provide additional web links explaining necessary notions or "theoretical medical background" in the right manner.

Case Three – Unmasking the Demagogy

Many statements of politicians sound attractive and plausible in situation we don't have certain information or data. After acquisition certain knowledge, the statements suddenly become fishy. For voters it would be very useful to have a tool which alerts him or her on suspicious parts of the text when reading for example a newspaper article. Shortly, to have an instant personal web assistant that helps with reading: "...look, this suggested action will probably lead to the growth of the state debt and today the debt is dangerously high, you didn't know, did you...?"

Similar functionality can be used in another context: in discussion forums below newspaper articles. Together with human comments there can be posted automatically generated comments concerning and commenting probable ignorance of the author and recommendate relevant data sources or opinions. It is a chance to work with people with extremist point of view. It is a group of people that should be informally educated – these people in fact are not faced with adequate responses to their extremist opinions and relevant arguments against simplified viewing of the world. Ignoring of extremist expressions can fetch us many social problems – only a few people are aware of potential danger of progressive growing number of extreme radicals.

Case Four – Intelligent Product Descriptions

Let us start with a particular example: a grandfather is trying to select the right present – a smartphone for his grandson. Typical product detail in a typical e-shop provides him a plenty of incomprehensible information. What does it mean that there is a 5Mpx camera? Is it much more worse then another smartphone with 12 Mpx camera? Is it an essential difference or is it something marginal? In reality, grandfather would probably go to a normal store and

ask the shop assistant for help. A honest shop assistant who wants to give him a serious help should at first find out the adequate level of communication and map grandfather's ignorance in this field and only then explain the foundational knowledge necessary for the right choice.

It is obvious that probably all actions in this process can be automatized. Analysis of grandfather's ignorance in the field of smart mobile devices can be used for preparing a personalized versions of a product detail pages which can include many explanatory notes showing what is important or not.

In general, implicit ignorance can be turned into a explicit knowledge about someone's ignorance and this knowledge can be used for better personalization. It is the way to the next level internet browsing, e.g. comfort e-shopping.

Ethical Issues

Presented examples of dark future naturally lead us to several ethical issues. How to avoid these negative consequences? How can we prepare the law system for sketched challenges?

One of the easiest ways is to leave *status quo*. This approach is based on the assumption that relevant parts of the law systems are general enough to cover all of these negative cases. Many instances of them can be handled simply as frauds. Analogous principles we can see for example the Act on Telecommunication, where many aspects of this Act were naturally extended to internet services etc.

The alternative way is to try to describe explicitly what is allowed to do with different pieces of knowledge about ignorance. This approach is analogous to the practical use of personal data protection acts. This way seems to be rather disputable. Today, there are many trials concerning personal data protection and we are not able to predict the result. The problem is that the borderline between correct and incorrect or unlawful manner is extremely fuzzy.

Another problem is that the law is one step (or more steps?!) behind technology. Some company tries to use some innovative approach, i.e. some application of datamining, then this action is found immoral and then, after that is such action covered by law. First company which tried that got a comparative advantage, but potential successor couldn't – is it fair or not? Similarly in different countries with different law systems.

In this situation it is good to have on the mind that **legislative background is only a tool for some purpose,** i. e. avoid examples in the dark section.

From my point of view the biggest safeguard is not the law system, but the openness of the online world. Although it will be easy to run fraudulent business based on using mass processing of data about ignorance, it will be also big opportunity for services protecting **against** these crimes. If someone tries to start up his business upon personalized “complicated terms and conditions“, there will surely be immediately a service which translates complicated text into the form understandable for the particular client, it will be also easy to create a recommendation system which warn about these unfair institutions or individuals.

There is also a less discussed topic concerning wide usage of machine learning classification tools. What can an individual or institution do, when the wrong classification has negative,

unfair consequences? Maybe there will be a trial with a team of data analysts which used certain method which caused big loss. Who has a bigger part of fault? The person who prepared the training data? The person who decided to use the particular method? Or the person who has specified the parameters of the model…? And who will prove this? When we are considering mass usage of datamining methods on data about ignorance, these questions becomes more and more important.

Bibliography

[1] Fagin, R., Halpern, J. Y., Moses Y., Vardi M. Y. *Reasoning about knowledge*. MIT Press, 2004

[2] Stephen Baker. *Numerati*. Computer Press, 2010.