Information Services & Use 36 (2016) 119–125
DOI 10.3233/ISU-160799
IOS Press

The data you have… Tomorrow's information business

Marjorie M.K. Hlava
President, Access Innovations, Inc., 4725 Indian School Road NE, Suite 100, Albuquerque, NM 87110, USA
E-mail: mhlava@accessinn.com

Abstract. How do you curate your data today to ensure you can capitalize on it to build a successful information business of tomorrow? Can artificial intelligence support and enhance human intelligence? What activities might help us build information services that are essentially the foundation of artificial intelligence in this information community? Where is the best source of user behavioral data? This paper will attempt to answer these questions and more regarding the status of artificial intelligence today.

Keywords: Artificial intelligence, machine learning, natural language processing, automated language processing, inference engines, text analytics, computational linguistics, word co-occurrence, Bayesian analysis, neural nets, vector spaces, automatic indexing, automatic translation

1. Introduction

Artificial Intelligence has long been the Holy Grail in information and computer science. How can we get the computer to think and act like a human? The computer is so ubiquitous in our lives today it is difficult to remember a time when we weren't computer-assisted in our daily lives. So where are we on this journey and how much can we expect to see in our life time?

I think of artificial intelligence (AI) as being on three levels: Narrow, General, and Super-Intelligence.

Artificial Narrow AI is also known as weak AI. It often covers a single area or domain. It can, for example, play chess – and very well. Narrow AI is used in service robotics and heavily implemented in machine robotics. It is quite pervasive in our lives today.

Artificial General Intelligence, which is also known as strong AI or Human Level AI, involves more reasoning, problem solving, and the creation of complex thoughts based on experience. These are the learning systems that are expected to perform any intellectual task that a human being could perform. At this level, the computer is thought of as being as smart as a human and is the main mantra of AI.

Artificial Super Intelligence is where the computer is smarter than a human, or at least it thinks it is. This is where the wisdom of the machine comes into play and allows the computer to implement solutions based on its knowledge. We would also expect this system to have the nuances of human interaction represented with appropriate accompanying social skills.

Most applications have gotten smaller and the everyday breakthroughs are incredible. In times past, we had to either need to ask a librarian to do a search or go to the reference shelves ourselves to look up an answer. After dinner with my family, we often covered the dining room table with reference books while several of us debated the historical facts supporting a statement.

0167-5265/16/$35.00 © 2016 – IOS Press and the authors.

2. Artificial Intelligence today

Now it seems that everyone just grabs their smart phone and asks Siri or her equivalent for answers. We ask "People" for facts such as how to get to places, where is that restaurant, where is the closest Costco, who was King of England in 1437, and more. This Artificial Intelligence system, which also translates the spoken query to an online search, parses the query, explodes the synonyms, executes the search, translates the answer back to voice that says "here is what I found on the web." I can ask it to call my husband, give me mobile directions to a restaurant, and a whole lot more. Sometimes I wonder "what did I do without it?" I had a glove box full of maps and an old phone book in the car to look things up. Where are those maps and that phone book now?

At the same time, some things are beginning to look different in our world today. Moore's law was busted in 2015 for the first time and technology is changing. Moore's Law said that technology change doubled every eighteen months. Now the hardware changes are not moving as quickly, but the software implementations are diverse and exploding. It often seems that games and gamification are leading the way. It's not just capacity anymore that makes the difference. It has been a meteoritic curve.

Other changes due to the breath of capability and hardware support have markedly changed as well. In February 2016, a one hundred year old theory was finally factually verified. Gravitational waves theorized by Einstein in 1919 are now an established fact. Other fun and amazing things from science fiction are regular working applications now, like the Star Trek Tricorder, a version of which can be used now in the field by geologists.

Our cell phones are no longer just mobile devices for making calls. They are smaller and much more capable. Mine is a phone, a computer, and a flashlight and holds my novels for a long trip! I can watch a movie, buy a latte, hail a ride, deposit checks, start the washing machine or check the security on my house. If offers so much more than Ma Bell offered using that black land line device only a few years ago.

AI is now used in warfare to provide facial recognition, programmable weaponized or reconnaissance drones, or mobile assassins. Robots have changed from block-headed TV screens on mechanical legs to storm troopers with Plexiglas monitors that acquire information.

A look back at Roman siege craft and works shows wooden structures, catapults, and trebuchets nearing castles with ramparts, towers, and moats. Now there are satellites transmitting over radio as well as wireless unmanned aerial vehicles (UAVs) and other client systems that can pinpoint information in real time. These systems have been moved to the private and commercial sectors as well, providing information on weather, traffic patterns, utility repair, event management, and sending packages via Amazon and FedEx at incredible speeds across the world.

One of the much-heralded recent AI events was the appearance of IBM Watson on the TV game show Jeopardy. Watson looks like a computer and beat the other contestants on Jeopardy. The stage view showed two human contestants and a computer screen for Watson. In fact, behind the curtain it looked quite different – the second room needed for Watson during the Jeopardy show held 53 IBM Mainframe computers! The real IBM Watson needed to perform on Jeopardy included:

- 90 IBM Power 750 Express Servers @ $34,500 each;
- 8 core processers per machine;
- Each with 4 sets of CPU's;
- 32 processors per box;
- Total of 2880 CPU's;

- Three second response time;
- Total processing power 80 teraflops;
- 1 trillion apps per second;
- $3 million just for the hardware.

In addition, there was the building of the Watson software, which included many databases, dictionaries, voice translation, speed of lookup and all the collateral parts.

All of the items I have mentioned so far and which I believe the reader will be well acquainted with are level one AI. What does a real reasoning system look like?

3. Current AI systems

Remember the computer Hal in 2001 Space Odyssey? He said to the remaining human, “Sorry Dave, I can't do that.” The computer, Hal, had reasoned that in fact it was superior to the human, and needed to survive without Dave the human. Another example is the recently revitalized Stepford Wives. They were not really human, but rather made to seem human and be the perfect wives for a group of men. Until a real human woman came into their midst many of the men had no idea that their perfect females were in fact computer-driven.

So how did we get from being thinking sentient creatures like the chimpanzee, to humans, to the ancient Roman trebuchet, to the current use of computers and artificial intelligence? Where are we on that journey? And more importantly to us today, what about the application of AI in information services and for publishers? How can we leverage that technology for better sales, discovery, and delivery of our products and services?

When I think of AI today I see the confluence of several forces. They come together to form a field that we can harness in many ways:

- Artificial Intelligence;
- Computational Linguistics;
- Automated Language Processing;
- Natural Language Processing;
- Co-occurrence;
- Inference Engines;
- Bayesian Analysis;
- Neural Nets;
- Vector Spaces;
- Text Analytics;
- Automatic Indexing;
- Automatic Translation.

All of these, just like the IBM Watson, work more accurately with a dictionary or taxonomy in use behind the scenes.

One example of an AI Application for publishing is in detecting SciGen or artificially-generated technical papers. Automatically generated “professional” papers have been around, embarrassingly, for a while. MIT students developed the system in 2005 and since then it has been broadly used to create papers in scientific disciplines. We built a program to detect these papers either in a large corpus or on author submission. We have found them *everywhere* we looked. These automatically-generated papers give peer review a black eye: they are easy to create, and really hard to detect. Although there is no coherent story line, the dictionary-generated papers use big words strung together so they look like real scientific papers. A peer reviewer reading such a paper after a long workday will often just pass the paper along as acceptable since it is totally obscure, but sounds really good. Here is a sample paper segment from a SciGen Abstract.

> *"Many theorists would agree that, had it not been for DNS, the simulation of randomized algorithms might never have occurred. In this position paper, we demonstrate the investigation of expert systems. In our research, we demonstrate not only that gigabit switches can be made modular, read-write, and constant-time, but that the same is true for online algorithms. We confirmed that while the seminal encrypted algorithm for the evaluation of RPCs by Takahashi is impossible, Internet QoS and operating systems can interfere to realize this ambition. Our framework for evaluating hierarchical databases is daringly encouraging. Similarly, we also introduced an analysis of the partition table. We disconfirmed that although simulated annealing can be made secure, multimodal, and pervasive, rasterization can be made ambimorphic, wireless, and signed."*

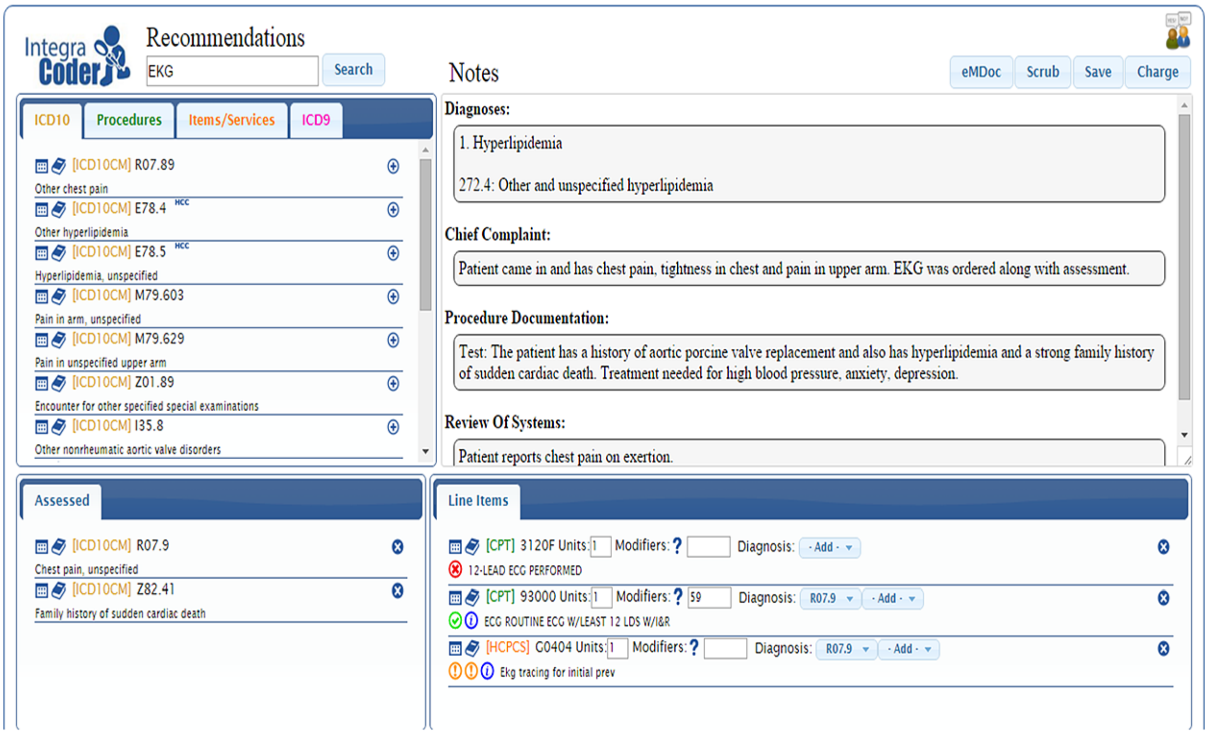
Look at the last, underlined sentence. Total gibberish. How do you detect these things? We do it using a combination of AI techniques in order to find the patterns.

We have available many different techniques and systems to enhance our publications, deliver them effectively, and capture original data to work on. These include, but are not limited to, logical techniques, digital reasoning, end-user attributes, procedural reasoning, machine reasoning, context attribute reasoning systems, automated reasoning, data attributes, conclusion generation, expert systems, gesture and emotion recognition, sentiment analysis, and I could go on, but they are all variations on a theme. They all depend on available knowledge. They do not create knowledge or wisdom.

Another area that has engaged in AI is medical applications. Medical coding and billing is one of them. The Center for Medical Systems finally won Congressional approval to implement the ICD-10 coding system in the US. The rest of the world adopted it years ago. The code set went from seventeen thousand codes to over one hundred and seventy-eight thousand codes (including hospital codes). The industry feared their own Y2K Event. The huge number of codes and the implementation of them could have had a horrendous effect on the cash flow of medical systems already burdened by heavy regulation and oversight. Some of the AI used in medicine is to provide help in the medical diagnosis itself. The patient encounter is the baseline for the information available via the Electronic Medical Record (see Figs 1 and 2).

4. Okay it looks good, but can you afford it?

Many of the systems that you see today have wonderful demos that start with the classification and mapping of application attributes. These are beautiful, but usually handcrafted, so they are not so much an artificial intelligence application, but rather a custom built imitation of one. There are computationally-intensive demos that are expensive to create, must be further tailored to your data, and create system lifelong maintenance with handcuffs to the vendor who built the system for you.



EKG

[ICD10CM] R07.89
Other chest pain

[ICD10CM] E78.4 HCC
Other hyperlipidemia

[ICD10CM] E78.5 HCC
Hyperlipidemia, unspecified

[ICD10CM] M79.603
Pain in arm, unspecified

[ICD10CM] M79.629
Pain in unspecified upper arm

[ICD10CM] Z01.89
Encounter for other specified special examinations

[ICD10CM] I35.8
Other nonrheumatic aortic valve disorders

Diagnoses:

1. Hyperlipidemia

272.4: Other and unspecified hyperlipidemia

Chief Complaint:

Patient came in and has chest pain, tightness in chest and pain in upper arm. EKG was ordered along with assessment.

Procedure Documentation:

Test: The patient has a history of aortic porcine valve replacement and also has hyperlipidemia and a strong family history of sudden cardiac death. Treatment needed for high blood pressure, anxiety, depression.

Review Of Systems:

Patient reports chest pain on exertion.

[ICD10CM] R07.9
Chest pain, unspecified

[ICD10CM] Z82.41
Family history of sudden cardiac death

[CPT] 3120F Units: 1 Modifiers: ? Diagnosis: - Add -
12-LEAD ECG PERFORMED

[CPT] 93000 Units: 1 Modifiers: ? 59 Diagnosis: R07.9 - Add -
ECG ROUTINE ECG W/LEAST 12 LDS W/I&R

[HCPCS] G0404 Units: 1 Modifiers: ? Diagnosis: R07.9 - Add -
Ekg tracing for initial prev

Fig. 1. IntegraCoder electronic health care analysis.

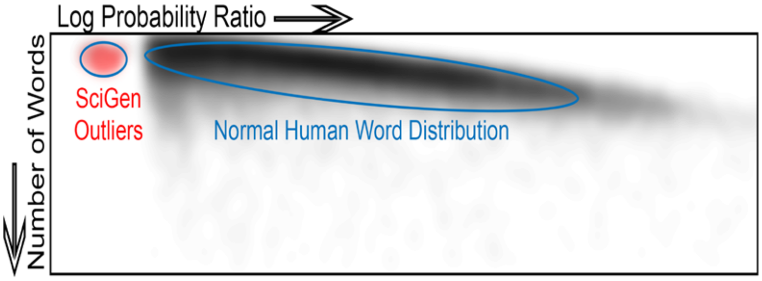


Fig. 2. SciGen detector probability distribution plot.

5. Enhancing content

What do you have that you can use now? AI today is algorithms – based on logic. It cannot make inferences without semantic control. Look at how an ontology statement might work:

<DOI1234> has topic <horses>
<DOI1235> has topic <horse>
<DOI1236> has topic <equines>

These are completely different sets and if they are not connected, they are statistically and therefore informationally worthless. The concepts themselves, represented by these words, are what we need to make into something we can interpret. Or, a machine can interpret it for us. The concepts are represented in many ways: pictures, videos, words. We use all of them and know what we mean, but the computer is very literal in its interpretation. The computer needs content and synonyms to make the proper judgment and return contextually-appropriate results in a search.

The metadata which is kept or separated by us into several elements like journal name, issue, date, DOI, paper, authors, author affiliations, date, price, etc., needs to be gathered into discrete fields or XML elements. Most publishers have done this. Then it is good to know the topics or subjects of each object (article, for example, or technical paper). In order to count it, we have to computer index, or sort it, and then it becomes part of the information process.

With those basic elements in place we can add logic statements, like triples, which can then feed the appropriate algorithms for everything from search to advanced analytics. But you can't mess up the units, or the analytics will be rotten. You can, however, make the most of the data collected that is your content by ensuring that the core data is correct and consistently managed. Other things added semantically such as concepts, topics, and taxonomies, make that data come alive in the AI sense. It allows inferences, finding meaning, and trends in the data that you already own.

Content has to go through the semantic "equalizer," or a controlled list of topics or taxonomy terms (a.k.a. ontology, thesaurus, subject headings, keywords, descriptors) which are needed to control the data to make it statistically accurate.

Using as many synonyms as you can find will also help leverage the data and increase discovery options for the users. Indexing rules need to be considered carefully by ensuring that the synonyms are applied, otherwise bubbles or groupings of like items are not covering the same material. For example, horse, horses, and equines are the same topic to us, but just a string of letters to a computer. Semantic normalization using synonyms and nested terms tells you how many people are really reading on any given subject, and how specifically they are interested in it. Your search logs will also provide the level of specificity the users of your data expect.

Coming back then to the basics for Level 1 AI, we find that the building blocks are largely things most publishers already have in varying amounts. They just are not tied together to truly leverage the data in meaningful ways. We are still too tied to the historic library applications and have not moved to the new frontiers of information science using AI. We have the concepts – the subject metadata – available for capture via topical offerings, search logs, and web navigations that are already held in classification systems and taxonomies or thesauri. Publishers can corral that information to support activities like automatic indexing supported by logical reasoning and accompanying rules created, which will allow tagging or mining the entire corpus published for trends, analysis, search, and discovery. Search and discovery are fully based on leveraging the word base used to tag the data – optimistically speaking – in a consistent, deep, specific, system using the term aliases. Once done, other pleasing applications for users like visualization and personalization of the user interaction can happen easily.

What about the other levels? Will we reach the Holy Grail? Do we have fully implemented artificial intelligence? Well, not yet. Will it happen in my lifetime? I give it a fifty-fifty chance. In the life of my grandchildren? I would say probably so. It is a brave new world with many unexplored frontiers for all of us. I look forward to the journey.

Acknowledgements

The author would like to thank Dr. Jay Ven Eman, CEO of Access Innovations, and Bob Kasenchak, Director of Business Development, Access Innovations, for their thoughts and contributions to this paper.

About the author

Marjorie M.K. Hlava is President of Access Innovations, Inc. Her research areas include furthering the productivity of content creation and the governance layer for information access through automated indexing, thesaurus development, taxonomy creation, natural language processing, machine translations, and machine aided indexing. She is past president of NFAIS (2002–2003), and presented the NFAIS Miles Conrad Lecture in 2014, and chaired the NFAIS Standards committee from 1999 to 2015.