Iuron.com Manifestation - Initial Reference (Draft)

Roy S. Schestowitz

Homepage: http://schestowitz.com E-mail address: r [at] schestowitz.com

12th October 2005

Version 1.0: a rough description intended to raise suggestions for further improvement, as well as identification of gaps and/or fallacies. Structure adheres to *streams of consciousness* at this early stage.

Abstract

This draft outlines a rudimentary plan for creating a knowledge engine that feeds on the World Wide Web. We begin with a survey of existing, relatively successful technologies, which are listed along with their shortcomings. We proceed to a succinct critique and end with proposal of a method that is tightly-correlated to semantics and relational attributes in text. Such information rarely, if ever, gets extracted *en masse* despite its enormous potential and assimilation to innate human rationale throughout exploration and learning.

1 Introduction

THE vocation of search technology – and in particular its *scope* – seems worryingly narrow-minded. We live in an age of certain complacency with what is available for information extraction and discovery. The common Web surfer has a much-unjustified expectation that whenever a *fortunate* query gets invoked, he/she will be referred to a page which leads to an answer. The way this is done far from ideal or even acceptable if state-of-the-art research is accounted for. The process of answer-seeking is subjective and overly time-consuming at present. It should be possible to retrieve answers at the speed of will (or speech). Referral to a human professional – a field expert that is – is still more fruitful and open-ended than the on-line scatter of pages.

The Internet as we know it is transforming as we speak. We begin to incorporate, intentionally or unintentionally, structural and relational data like XFN, document classes and closed networks of collaborated knowledge – a finite, closed-ended universe of manageable and interpretable facts, that is. Further informative text at code-level is presently embedded which assists in fetching semantics, thus optimising exploration and encouraging cross-site collaboration.

With Web 2.0 (as it is commonly referred to) on the horizon, everything migrates to on-line storage units (private or public) so that it resides coherently and cohesively in a single domain. Having got huge heaps of knowledge assembled and inter-linked, traditional search engines would proceed to scanning pages and extracting key words from them to form indices. This is a most fundamental – some would say primitive – way of reflecting on page content, which is crucial when one undertakes a most laborious and error-prone task. That task involves covering billions of pages from possibly questionable sources, which are written in different languages and comply with different contexts.

Word density, word proximity and the like are

currently analysed by market leaders, but no actual knowledge is formed. The potential for forming hypotheses and testing them is missed entirely, even discarded. Words are treated as atomic elements within a large pool and are perceived as merely unrelated entities despite the fact that, in the mind of the author, a continuous flow of thought was stirring.

When a page gets composed, the final outcome is a document where an actual story is told so arguments are provided in a logical order and each argument is related to its neighbours. Missing that observation makes an algorithm deem to weaknesses, if not utter failure.

2 MOTIVATION

Given all the data that is contained in our E-mail box(es), our files (photos, documents and sounds to name a few), would inference not be one natural direction to follow? Extracting the semantics from our data and forming a network of knowledge will enable us to search for answers rather than arbitrary pages that approximately resemble our query text. Taking personal search (limited scope due to privacy) as an instance, the implications can be particularly revolutionary. Having got large lumps of data, we should be able to perform a query, using natural language, to receive plenty of information about a person (or place) including relationships, photos, etc. See the short discussion on privacy controversies in §B.

3 THE PROPOSED ALTERNATIVE

Rather than "search engines" we ought to be talking about "knowledge engines", or at least that is my contention. Googlism is worth citing as an initiative that took a similar approach, albeit it became static and was bound to a *search* engine that indexed rather than learned. Googlism used indexing as a 'bridge' to the formation of knowledge. Plenty of data was already drained or diluted by the time it piped through to the basic learning method.

There is currently no barrier that stands in the face of implementing large-scale knowledge-bases apart from computer power, bandwidth (for data quantity) and amount of code with intricate dependencies. With Open Source software (making use of public and Open sites, e.g. Freshmeat and Sourceforge), prototypes ought to be achievable within a realistic timescale.

Imagine a neural network out there which rather than contain text with your name has got complex knowledge about who you are. Moreover, it can answer questions that involve you and may eventually become too complex to be explored or administered by a human. To many bodies including governments this would be invaluable. Real incentive and desire is certainly there, yet privacy can be jeopardised, as always (c/f §B). Iuron aspires to be a beginning – a seminal manifesto – of knowledge engines as opposed to search engines.

4 FUNDING AND SUPPORT

The entry barrier for a project such as this is extremely high. In order to just test an algorithm, vast amounts of data need to be analysed. The more data is available, the better the results should be. Iuron aims at funding from University incubation or possible funding from investors and/or relatives. Here are a few possible funding frameworks, listed in order of precedence or likelihood of success.

- ❖ Bio-informatics or The Information Management Group (Taylor, Rector)
- ❖ Automated reasoning (possibly under the umbilical cord of 2^{nd} Ph.D. program)
- ❖ University incubation
- ❖ Investments from outside
- ❖ Google support (DiBona ties: code and Open Source)

Experimentation can begin at a small scale by being used temporarily on the University site[1].

[1] Inspiration here stems from a past success story, much like the early days of Google at Standford University. The company's

It is worth re-iterating and stressing the fact that Iuron aspires to be somewhat of a knowledge base, hence it must exploit an abundance of information and sophisticated machine learning approaches rather than simple word matching, counting, and storage. Consequently, it needs high funds to become viable or even be possible to argue about as successful (and verging a point of empirical positive evidence). Due to scale that is required to make our statistical sample for learning sufficient, ad-hoc methods must be devised, particularly at the start.

We may take a genetic algorithms approach, whereby weak facts are discouraged and repetitions are perceived as encouraging. Larger scale will lead to more accuracy, saturation and reliability. This method may also be rather immune to spam unlike some traditional searches, but PageRank-like mechanisms still need to be put in place.

It is important to consider conflicting interest and deceiving knowledge sources that use repeatable false content ("spam" where its meaning become "mass lies"). For example, a pharmaceutical company will have financial incentive in spreading a false word, according to which their drug is the best solution to an illness of some kind.

Trust is extremely important, much like attempts at TrustRank and human moderation at DMOZ (a non-profit Web directory) have shown. DMOZ gets a positive mention as opposed to Yahoo's corporate-inclined directory where money warrants listing. Another problem with PageRank is that ranks can be purchased in the form of link. So, power and influence can be *bought* rather than rightfully *earned*. Due to the scale of the Internet, fraud cannot be controlled by a human. The system must be self-sustaining. Algorithm secrecy and obscurity is often the vendor's solution to the issue. Comment spam, "Googlebombing" and the like are some of the detrimental by-products of a deficient algorithm.

These challenges or barriers cannot be *trivially* solved by this paradigm which is knowledge engine. However, its extra complexity should open more doors to optimisation, refinement, and improvements.

founders consumed almost half the bandwidth on campus while expanding.

5 THE APPROACH

At a rather shallow level, the approach can be outlined as listed below:

- ❖ Strip pages from tags (further help is available in the newsgroups)
- ❖ Use headers and tags to highlight important facts
- ❖ Identify synonyms (language becomes an issue, but maybe translation can bridge the gap)
- ❖ Collect a list of facts from the page/s in question

For data reliability, we may consider using Wikipedia as a better facts source where mutual moderation is perpetually forced. The grand scheme is to crawl pages and not to index and summarise them, but rather to accumulate knowledge, much like a human reader would do. It is always worth remembering (caching) sources of information to refer the reader back to. This would establish confidence and further breadth for the user's mind. Better priority should be given to pages with stronger PageRank et al., i.e. pages with more inbound links. Moreover, it is worth using age of domains, professional affiliation and so forth as factors; all of these are also worth scoring accumulatively. *Impact* should be emphasised as an important aspect in oder to avoid false facts from ever being absorbed as truthful ones.

As for the user's side, voting mechanism can be used by the engines or even explicit queries made in natural language and then interpreted logically (first order predicates). For example, the user can ask a question or provide some query terms. He/she will consequently get answers sorted by certainty of response/answer with relevant links/pointer to the sources; snippets as well can be attached to answers if cache is available to access.

6 FEASIBILITY

Neural networks are not quite so feasible as they are too many parameters (hence dimensions) to consider. Each word is a dimension of its own, but it is worth understanding how search engines overcome that same problem.

The need for caching is apparent, yet in our case it does not involve indexing. The cache is in some sense the 'brain' of the engine. Apart from knowledge indexing, this cache may be needed for providing references along with an answer to a user's query (which should ideally be labelled a "question").

To test the idea at a very small scale, experimental engine can be made available for `schestowitz.com` (in case of University refusal) and a little search bar added to the main page.

7 PITFALLS

Search engines and machine learning algorithms are poor at detecting and understanding social patterns including sarcasm and humour. A highly-cited page can in fact filled with humour that would be misleading to a naïve engine. For example, with popular phrases like "when pigs fly", it can be inferred that "pig" is a form of bird because birds fly. This winds up forming a wrong taxonomy, whereby pigs have wings and a beak.

8 OUTLOOK

If queries are ever made using natural language and require no further user intervention, voice recognition and vocal output is worthy of imagining. Visualise a scenario where you ask a question and get a series of answers with level of certainty for each. If you are unsatisfied with the foremost (top-level) response, you may explore the other possibilities which have been perceived as reasonable by the knowledge engine. You can also explore laterally, forming your own opinion based on the facts which are available in the form of strongest citations.

Privacy become a major issue too. One must separate out names of individuals and businesses or have it tightly moderated, which is bad practice. Censorship by service providers over content is often harshly criticised, in particular Google's political censorship in China.

9 SUMMARY

I have described a family of valid methods for forming knowledge based on large amounts of public or personal data. These have the potential of resolving many issues such as purchasing of impact and influence in the form of links, thus achieving precedence. Moreover, the process of so-called answer extraction is made quicker, simpler, more precise and more related to the way our human mind works, namely learning and understanding rather than accumulating text.

APPENDICES

A DOMAIN AND PROJECT NAME

The domain name was a compromise given what remained available according to ICANN.

`Nueron.com` or `Nuero.com` would have been somewhat ideal had they been available, `meuron.com` (micro-neuron) has been taken by a Japanese already, `Euron.com` was at some stage selected, but it turned out to have been occupied too. `Iuron.com` was a short and elegant name that did not contain the geographically limiting 'euro' either. The preceding I refers to 'Internet' which is the main target domain.

The reasoning behind the name is associated with aspiration to form proper knowledge with a networked hierarchy. Time will tell if the complex task can ever actualise without ad-had workarounds and simplifications.

B Applicability

Further on the issue of privacy, it is worth subdividing the possible application of knowledge engines.

Private scope manage one's own knowledge used as an aide for memory (a mnemonic or 'life manager').

Global scope share knowledge that does not involve individuals and bodies. There is a fuzzy seamless border to consider here, but indexing of scientific knowledge, for instance, can greatly benefit from knowledge bases as science is less subjective than humanitarian matters.