Text mining essay

Vaim Faqs Steve Kimbrough May 10, 2006 1 What does the word vaim mean? The word vaim has two meanings: 1. In Estonian, the word vaim means ghost (http://et. wikipedia. org/wiki/Vaim). See http://www.

logosdictionary. com/pls/dictionary/new dictionary. gdic. sl? phrase code= 5573701 for the pronunciation.

2. The word vaim is an acronym, abbreviating Value-Added Information Mash. The two meanings are unrelated. It is permitted to pronounce both words in the Estonian fashion. 2 What is a mash? From http://en.

wikipedia. rg/wiki/Mashup %28web application hybrid%29: A mashup is a website or web application that seamlessly combines content from more than one source into an integrated experience. And The etymology of this term almost certainly derives from its similar use in pop music where DJ’s take the vocal track from one song and combine it with the instrumental track of another song resulting in an entirely new composition. In the lingo, Web mashing results in a Web mash or mashup. See the Wikipedia article for examples and further information. Programmable Web (http://www.

rogrammableweb. com/) is a Web site devoted to Web mashing. 1 3 What is an information mash? The original (or at least an early) reference appeared in a blog by Ellen Miller of the Sunlight Foundation (www. sunlightfoundation.

com) on April 28, 2006. In her blog (http://www. sunlightfoundation. com/node/465) she writes: Information Mashing. Don’t you just love that term? It’s one of the ma jor goals of Sunlight and while we’ve been working

on it for the past couple of months we have a ways to go before it happens in any substantial way.

Our goal is simple: integrate in a user-friendly way individual data sets (like campaign contributions, lobbyists and government contracts) that makes the whole larger than the sum of its parts. We'd like to create something we've dubbed an " Accountability Matrix. " A website where, with one click you can look up a ma jor donor and see not just their campaign contributions, but also their lobbying expenditures, the names of members who've ? own on their private jet, the names of former congressional sta? ers they've hired, and so on. In a nutshell, we want to make information more liquid and more accessible to the public. Although the information mashing she writes about is broadly on the sub ject of politics and current events, the concept of information mashing is not so restricted.

An information mash is any sub ject-focused aggregation of information from multiple sources that achieves the-whole-is-larger-than-the-sum-of-its-parts status. In other words, meaningful, useful, non-trivial integration of information from several sources. What is a value-added information mash? Value-added implies the presence of a signi? cant additional element of information process- ing, indexing, categorization, and so on. Information is not only collected and aggregated, but new information is added, typically through indexing, association of items in the dif- ferent aggregates, and other processing. The information masher may also add original information, not available from other sources.

Value added will often come from employment of advanced software technologies. Ex- amples include language translation, information extraction [JM02], associative indexing and retrieval [LD97], word pattern visualization [DKP00], data mining techniques, textmining techniques [WIZD05], literature-based discovery (aka: knowledge discovery) tech- niques [GLF02], faceted classi? cation (http://www. kmconnection. com/DOC100100.

htm), concordances, and others, such as [BK02]. 2 5 Could you be just a little more speci? c about uses and applications? Yes, just a little. Detailed discussion is apt in other venues. Perhaps the key idea is association. An information mash facilitates ? nding signi? cant associations (or signi? cant lack of association) among information items (data, documents, etc.

). Investigators seek more than lists of relevant records. They seek patterns of infor- mation evidenced by associations among information items. (See [DKP00] for discussion of record-oriented versus pattern-oriented information retrieval.) Here is the sort of example that the Sunlight Foundation (above) is interested in for information mashing.

Company A sells an item of real estate to company B, which then quickly sells the property at a large pro? t to company C. However, both company A and company C are associated with company D, which is a defense contractor recently receiving a large contract. And company B is owned in part by a Congressman with strong DoD links. Similar events led to the conviction and jailing of Congressman ‘ Duke’ Cunningham, and he is hardly unique in the history of discovered crime. Cunningham’s crime was detected,

or at least hypothesized, by publicly available in- formation. The property sale records are public information.

The public contracting records of the companies are public information. Much information is available about the various companies, and a great deal of information is available about public ? g- ures such as Congressmen. An information mash can pull together information from multiple sources and assist users in exploring for evidence of interesting associations. Here are commercial examples. • IP (intellectual property) placing.

A ? rm has patented a kind of plastic that is made entirely from renewable re- sources (cf. http://news. bc. co. uk/2/hi/science/nature/4191737.

stm), and is interested in licensing the IP. In searching for potential partners or uses of the IP, the ? rm will use descriptors of the IP (e. g. , ‘ biodegradable’, ‘ water resistant’, ‘ made from biological materials’, ‘ used in kitchenware’, ‘ orange peels’) to locate relevant infor- mation. This information will be contained in a wide range of documents, including patents, patent applications, SEC ? lings, newspaper reports, corporate annual re- ports, general Web documents, internal technical and marketing studies, and so on.

IP placing is a fundamentally creative process and cannot be fully automated. An information mash can, however, assist the process by producing information rapidly and on demand, searching in a focused manner across all relevant sources. 3 • Investment analysis. The information needs for investment analysis are essentially open-ended. Any in- formation

with predictive value is relevant, but what that information is is not fully known.

The information sources that are known to be useful are many and disparate, including: – Market data on ? nancial instruments (stocks, bonds, etc. – Regulatory ? lings – Patent ? lings – Annual reports and other company-generated data and documents – General news stories – Index or rating information created by third parties, for example sustainability ratings from sustainability standards organizations such as Social Accountabil- ity International (http://www. sa-intl. org/) and the ISO (http://www.

iso. org/iso/en/ISOOnline. frontpage). An information mash, and especially a more powerful vaim, o? ers the attractive prospect of material support for locating otherwise undetected opportunities and risks. Is this really new? Yes and no. As expansively envisioned here, few if any vaims exist today.

Yet there is much research, development, and deployment that is closely related at the least. Web mashups are appearing daily (it seems); it is an active area. Perhaps the most closely related established concept is that of a digital library (see http://www. dlib.

org/). And it is more than an analogy to think of a vaim (or information mash) as an electronic form of publishing, on the model of the Annenbergs' TV Guide and (earlier) racing forms. This is publishing that aggregates information and adds value to the package. 7 What next? Building, testing, using. In going forward there are four key issues.

1. What is the envisioned use and what value would it deliver? What is the topic for focus of attention? Who are the users? Why should they want to use the system? In short, what is the application theory—the working hypothesis for what the value proposition is—for such a system? See examples above. How broadly the vaim concept will apply and succeed is an entirely open question. 4 2. What is the information to be aggregated? Does it have valuable content? Is it available and usable for this purpose? 3.

What recovery techniques will be used? How will value be added beyond merely aggregating the information? How will associations between and among the collected information items be established and presented? 4. How can costs be recovered? Social activists (see the Sunlight Foundation, above), fanciers (people with an avid interest in art, sports, cats, etc.), and researchers may have incentive and capacity to provide the labor and know-how needed to build and maintain a vaim. For these individuals, publication and renown may su? ce, especially with support from funding agencies, federal or private-sector.

How might pro? ts be generated? Three main business models have appeared: (a) Sell advertising (b) Sell subscriptions (c) Open-source: make the system freely available; sell services related to using it. References [BK02] David C. Blair and Steven O. Kimbrough, Exemplary documents: a foundation for information retrieval design, Information Processing and Management 38 (2002), no. 3, 363–379. [DKP00] Garett O.

Dworman, Steven O. Kimbrough, and Chuck Patch, On pattern- directed search of archives and col lections, Journal of the American Society for

Information Science 51 (2000), no. 1, 14–23. [GLF02] Michael Gordon, Robert K. Lindsay, and Weiguo Fan, Literature-based discovery on the World Wide Web, ACM Transactions on Internet Technology 2 (2002), no. 4, 261–275.

[JM02] Peter Jackson and Isabelle Moulinier, Natural language processing for online ap- plications: Text retrieval, extraction and categorization, John Benjamins Pub- lishing Company, Amsterdam, The Netherlands and Philadelphia, USA, 2002. LD97] Thomas K. Landauer and Susan T. Dumais, A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychological Review 104 (1997), no.

2, 211–240. 5 [WIZD05] Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, and Fred J. Damerau, Text minging: Predictive methods for analyzing unstructured information, Springer, New York, NY, 2005. $Id: vaim-faqs.

tex, v 1. 6 2006/05/10 21: 45: 07 sok Exp $ 6