Querying and Personalizing the Web: A Multimedia Personal Assistant

Stephan Vock and Raphael Ochsenbein

Department of Informatics
University of Zurich
Binzmühlestrasse 14
8050 Zurich, Switzerland

Abstract. Searching the web and finding the information needed becomes more and more an almost impossible task. In the last couple of years we've seen new approaches to the web, such as the Wolfram search engine or some rather big adjustments, as seen on Google search. With K-Dime a personal image search filter there might be a new and even more sophisticated way to browse the web than we knew so far. By using Kansei Engineering to create a personal search engine for every user we might be able to improve search results and to decrease the time used for any particular search.

Keywords: Kansei, K-Dime, MOUE, MIKE, user model, image filter, image search, web, personalize, emotion, mood, sensitivity

1 Introduction

The internet is a seemingly unlimited, vast accumulation of information. For an unassisted human, it's easier to find a needle in a haystack than to find something specific in the internet. This has led to the current situation of the software giant Google, who started as search engine. They now provide not only searches but they also analyse content, act as intermediates for advertisements and have worked to create a popular smartphone operating system.
All of this, just because they saw the importance of connecting internet users with the content they were looking for.

We have read three research papers: QP, ME and BK. Those three introduce a new dimension to the quest of helping users to information. Normally a user could start an image search with the key-word "beach". Now QP tries to enable the use of subjective criteria to assist the human decision making. They envision a search not only for "beach", but a search for a "romantic" beach, whether that is a sunset or a starry sky for the particular human who ventured that search. Now subjective criteria do not only apply to images, but also to many different forms of digital information like music, videos or texts. While the idea is that in the future those media could also be categorized according to affective criteria, the K-Dime framework as it is introduced by QP is built for image queries.

MOUE (Model of User Emotions), introduced in ME, aims to create a model of user emotions. Therefore facial expressions are analyzed to evaluate the user’s current emotional state.

BK shows that through the use of fuzzy logic, the effort of creating such a personalized user model can be reduced.

In this essay we will discuss the advantages as well as the disadvantages of using a Kansei user model to assist human-machine interaction, especially in the area of image search in the internet.

To accomplish that, we will have to begin with an introduction of what the underlying design paradigm, namely Kansei Engineering, means. When that has been made clear, we'll sum up the architecture of the K-Dime. We will then contrast the K-Dime to MOUE, and add the proposal of BK. With that we will be able to evaluate the three papers as well as draw our own conclusion.

2 Main

2.1 Kansei Engineering

As most of you might not be familiar with Kansei Engineering we think, that it's beneficial to briefly introduce you to its background and main principles before tackling the K-Dime. The Japanese word 感性 (Kansei), which would be translated by sensitivity or sensitiveness, is usually interpreted as emotional or affective.

We find its origins in the early 1970ies in Japan. The term Kansei Engineering was mentioned for the first time by K. Yamamoto the manager of Mazda Motors in 1986. Since then, the early beginnings of Kansei Engineering, it became quite a big field of research, especially in Asia. As for Europe and America, Kansei was adapted only in the 1990ies but never became as important and popular as it was in Japan and South Korea (Schütte).

So what exactly is Kansei Engineering? Basically it is a design principle applicable in various fields of research. Nagamachi defines Kansei as "the impression somebody gets from a certain artifact, environment or situation using all her senses of sight, hearing, feeling, smell, taste as well as their recognition" [Schütte]. The goal is, to capture human feelings or emotions and to improve the human-computer interaction, as well as to level it with the daily human-human interaction we know from everyday life.

That means that Kansei is a process in which the AI is in constant contact with the outer world (the user). It receives information, processes them and communicates back to the user (ME: 49).

As Bianchi-Berthouze mentions, every experience for a human has two levels, a subjective and an objective level. The objective part is the easy part to define. For example, a human is a human for everyone. It doesn't matter if he is large, big or tiny; everyone has the same definition of a human being. With a correct definition, any observer can discern if an object belongs to the category "human" or not. More difficult are the subjective categories, because not everyone has the same idea of a good-looking human. Everyone has different feelings attached to that term. We can measure those feelings on three different levels: affection, mood and emotion (ME: 51). And even though it is hard to define words like "red" or "nice" we are able to construct prototypes for such words as we will see later on (Bianchi 49-53). Or how Schütte describes it, the main principles of Kansei Engineering are to identify the objects emotional properties, to find correlations between those properties and to design characteristics to describe those objects.

There are several ways to measure the Kansei, for example people's behaviours and actions, words (spoken and written), facial and body expressions and physiological responses (Nagamachi, 2001). Two of them, words and facial expression will be discussed later on.

Kansei Engineering is not only used in software engineering, but also in product design and marketing. In that case, products are designed to look tasty or healthy.

2.2 K-Dime

Now that we've elaborated what the Kansei paradigm means in the context of software engineering, we can finally ask the question: Why would we wish to use such an agent for image search?

As for the answer, imagine yourself wishing to send a friend a birthday card. You open a word processor and you type a witty remark about age and wish your friend all the best. After choosing a font that supports your message, probably something along the lines of Comic Sans, you decide to add some eye-candy. Therefore you switch to Google and type something like "happy birthday". Now you get several million results and you have the tedious task of filtering all the unsuitable images. You will probably need to browse several result pages to find an image that suits your text and friend. Wouldn't it be nice if you could just describe the feeling your image should convey, let's say "witty and happy" and then getting images that match your understanding of that? "Nice" is probably an understatement.

Finding images according to such criteria is what QP try to achieve with the K-Dime (see QP: 678).

With the framework "K-Dime" they aim to attain four goals they consider important:

Firstly they intend to support the user decision process by giving him the ability to select a given media by using affective and subjective criteria (ME: 103). They state that in a further step, the K-Dime should be extendable to support any possible digital media but at the moment, it's built to search for images (QP: 678f). Because the K-Dime can "interpret" multimedia (QP: 678), a "more natural" human-computer interaction can be attained (ME: 49). The K-Dime is capable of showing the user more relevant images than the currently available search engines, reducing the time and effort spent to find an image (ME: 78).

To do this, they mention as a second point, that "tailoring information to single users" is necessary (QP: 678). The K-Dime adapts itself to his user by creating a cognitive model of the user (ibid. 679). BK et al show, that because users have different physical interpretations of emotional words, adapting the Kansei user model to single users is of utmost importance (362). VE et al draw their inspiration from "human-human communication" (981). A component called MIKE was designed to interact with the user in a text based interface (ME: 73).

The last two points consist of reusing already available tools (e.g. Google and AltaVista image searches) and managing the multimedia information extracted from those tools (QP: 678). Since the algorithm for evaluating images is quite processing intensive (ME: 79), the automated image filtering with K-Dime allows browsing more relevant images through offloading some image evaluation functions from the human to the machine. In the conclusion of our essay we will discuss the advantages and disadvantages that are a product of this approach.

Keeping these four goals in mind, we can now describe the implementation of the K-Dime. Actually the texts we quote range from 1999 (VE) to 2003 (MW), and over those 4 years some changes have been made to the architecture of the K-Dime. However the gist of it remains the same. There might be some inconsistency with the terms we use in our essay compared to our sources, but to facilitate reading we'll stick to the terms used by QP.

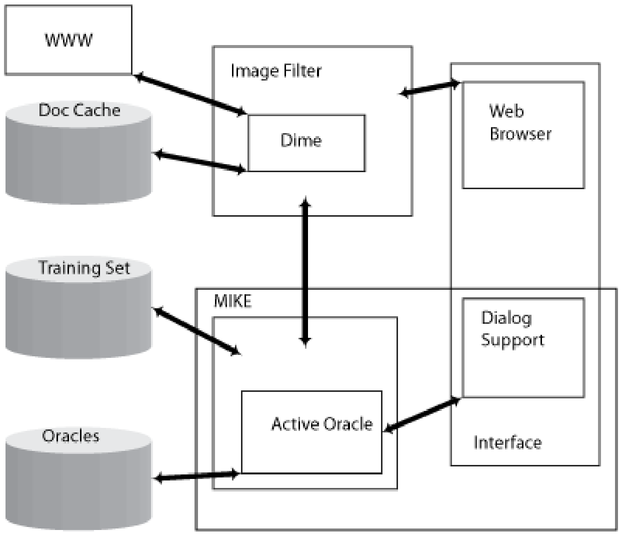


Fig. 1.

The K-Dime consists of three main modules: The Image Filter, the Oracles (or Kansei Agents) and MIKE (QP: 679).

The Image Filter is accessed by the user through a web interface. The user specifies "objective" search parameters which are used to query existing search engines (QP 680). The Image Filter retrieves the search results, analyses them and sends them to the active Oracle (QP: 680).

The image signature is used by the Oracles to classify the image in visual categories and to create "associations between perceptual states and impression words" (ME 71).

Oracles (or K-Agents) are the most important part of the K-Dime as they are responsible for the interaction with the user and subjective concepts (QP: 679). An oracle has four subcomponents: a user profile, an image-processing kernel, a learning kernel and an action evaluator (ibid. 679).

The user profile is used to identify the attributes defining the person who created one specific oracle (ibid. 679). It contains data like nationality, gender, age, job, and so on. But most importantly it also contains a list of known words to the user and their relation to each other (ibid. 679). These attributes are used as a fall-back option, if the active oracle does not know a word the user typed. In that case it looks for the word within the oracles of users with a similar profile (ibid.: 681). Therefore a database containing all known words and all oracles is also maintained (VE: 983). This mitigates the problem that each user has to define every word he wishes to use in a search.

Bianchi et al (ME: 69) make the assumption, that certain objective characteristics of an image evoke the subjective feelings an observer could see. The image processing kernel is used to isolate those characteristics (QP: 679). It simulates the way humans perceive images as follows: It first looks for an especially informative fixation point and creates a decrease in resolution from that point to the periphery to have a "retinal image" as seen by an eye (ME: 70). Then the images are sent to two distinct modules. The first collects information about colour distribution, brightness, edge orientation, shape extraction, homogeneity and contrast in a HSB (Hue, Saturation Brightness) colour model (ME: 70). This information is then used to create four signatures for each image: Colour, tone, shape and texture. The second module extracts the "context edges" at the fixation point to find the next focal point to create a new retinal image that can be analysed (ME: 70).

The learning kernel is responsible for learning the associations between "impression words" and the signatures the image processing kernel created (QP: 679). Every word known to the oracle is dealt by a "word module". The word module consists of four neural networks which learn the connections between the word and the signatures relevant to that word (ME: 71). Additionally the neural networks are weighted, as an example "fresh" could be linked to colours and "imposing" would probably be more linked to the shape of the object (ibid.: 72). The word modules are linked together in a graph to be able to connect related words. The possible relations are: "opposite", "synonym", "related" (ibid. 72). Additionally there are some predefined visual categories (they can be found in the table in ME 71, an example are "Red", "Orange", "Green" and so on for the category of Hue). Those words are also full-fledged word modules as their definition can be altered by the K-Agent and the User (ibid. 72). Their purpose is to be connected to other words with special "check-for" relations to enable the evaluation of the presence or absence of certain low-level features for other words (ibid. 72). To determine if a word is active, the K-Agent utilizes a threshold to check if there's a significant overlap between the signatures of an image and the values saved in the word module (ibid. 73).

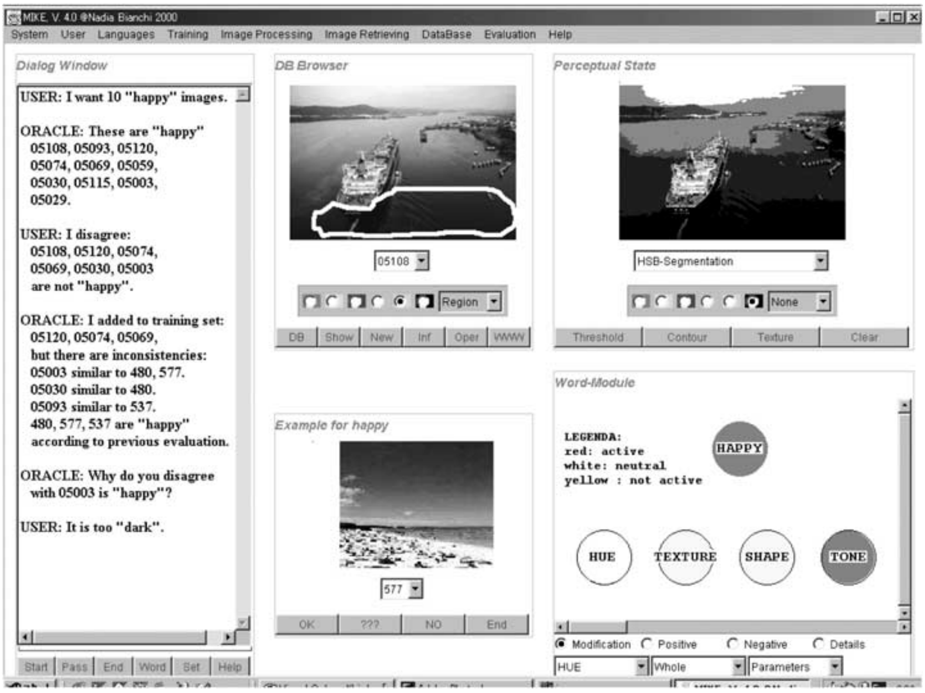
As for the technical specifications of the neural networks, I'd just like to quote the original article at ME: 72:

"The neural networks are of the three-layers forward type, trained by backpropagation (Rumelhart et al., 1986) with momentum. In our experiments, the input and output of each network are real numbers between -1.0 and 1.0. The learning rate is set to 0.01 at the beginning of the learning and automatically decreased down to 1.0E-6. The Epoch is 5000 and the threshold error is 0.001. These values have been determined experimentally. Unlike in typical applications where learning sets are constructed off-line, we are dealing with a continuously evolving training set (a by-product of the user feedback). Since the back-propagation algorithm usually performs poorly at such task (catastrophic loss of memory for example), an analysis of the structure of each module and the subsequent learning set is performed [...]."

The job of the action evaluator is to process the user feedback (QP: 679). The feedback is given through "interactive training sessions" and the action evaluator is able to use them to modify the training set and the structure of the word modules (ibid. 679). The environment for the user to oracle interaction is provided by MIKE, the "Multimedia Interactive Environment for Kansei Communication" (QP: 680).

In contrast to the Oracle who is in care for the subjective concepts of a user, MIKE mediates the interactions between user and Oracle (QP: 680). At ME 73, it is stated that three types of interaction are implemented: "Adding a new Symbol", "Reinforcing Learning" and "Symbol Specialization" (ME: 74). They are triggered by user input (in natural language), a visual example or self-externalisation (ibid. 73). The interaction is based on dialogues in "pseudo-natural language". It uses predefined visual categories (also found in the table at ME 71) and the words the K-agent already knows up to date (ibid. 73).

To illustrate how the interface of MIKE looks like, I'd like to use the screenshot from ME 74:



USER: I want 10 "happy" images.

ORACLE: These are "happy"
05108, 05093, 05120,
05074, 05069, 05059,
05030, 05115, 05003,
05029.

USER: I disagree:
05108, 05120, 05074,
05069, 05030, 05003
are not "happy".

ORACLE: I added to training set:
05120, 05074, 05069,
but there are inconsistencies:
05003 similar to 480, 577.
05030 similar to 480.
05093 similar to 537.
480, 577, 537 are "happy"
according to previous evaluation.

ORACLE: Why do you disagree
with 05003 is "happy"?

USER: It is too "dark".

LEGENDA:
red: active
white: neutral
yellow : not active

Fig. 2.

As we can see, the interface has different parts serving their task. Probably the most important one is at the left of the screen, the dialogue interface. There we can see a sample discussion between user and agent. The image shown in the DB-Window is referred to as "this image" and it's selected by the user through browsing the database (ME: 73). Below the DB-browser there is a dynamic interface, which is used to display a set of images (for example the output of a query). The agent also provides a tool to help with the externalisation of the neural networks for the user (ME: 75). The interface to help with that process is found on the right hand on the screen. With that interface the user can visualize the internal processes and the network the agent created for a certain word module (according to ME 75). Bianchi-Berthouze has the opinion, that visualized cognitive pathways can help users of K-Dime to detect inconsistencies in judgment and detect similarities (MW: 106).

2.3 MOUE

Bianchi-Berthouze presents us a second experimental framework for Kansei Engineering called MOUE (Model of User Emotions). Similar to K-Dime it tries to capture human emotions but in a very different way. So let us explain first how exactly MOUE works: A webcam takes a picture of the user. This picture will be analysed by facial recognition software and with the data gained, MOUE tries to label the user's current emotion. From now on everything is really similar to K-Dime. In a neuronal network every so far known emotion is saved and considered as a collection of emotion components or attributes as for example: valence, intensity and facial expression. Additionally they used culturally independent meta-semantic definitions of emotion concepts to describe the emotion and to communicate with the user about an emotion. After the picture has been taken MOUE tries to find the closest match found in the network and presents it to the user. He decides if MOUE's interpretation was right or wrong. If it was wrong but the emotion already known to MOUE then the attributes will be adjusted. If the emotion was unknown so far then MOUE creates a new node for the newly learned emotion connected to similar emotions with similar emotion components.

The idea is to create computational schemata of emotion concepts which can be extended by capturing other ways of human input like the sound of the voice or heartbeat with the goal to implement MOUE into different applications. For example into a music player which will pick songs according to your current state of emotions and if we look back at the beginning where we mentioned searching the web. With MOUE it would be possible to get direct user feedback while he's looking through information. If he feels lost MOUE could send him information.

The big difference between MOUE and K-Dime is now that K-Dime creates for every user a set of interpretation of words. Those interpretations are permanent and do not include the actual user mood. Users might have a different idea of a sunny beach when they're happy or when they're sad. And this is really the key feature of MOUE and even in its experimental state MOUE might find its way into actual products.

2.4 UNGAR

According to Kovacs et al one of the main problems with many Kansei User agents is the way how they create a user model for new users. Most of them just have offline generated user models and they just apply the one's fitting best to every new user. This may and most likely will lead to incoherence during the human computer interaction. These user models may approach the users own Kansei in some parts but are still far away from covering all the users ideas. What's missing is a personal and adjustable user model.

K-Dime handles user agents already in a different but still not perfect way. As we've seen every new user gets its own user model. The user just has to give K-Dime some personal data. Basing on that user profile K-Dime is looking for an already existing user with the most similar profile and copies its user model to the new user. Even though every user got now its own user model there is still a problem. K-Dime can't assure that every 25-year-old Caucasian men living in Zurich with a Bachelor in Physics shares the same interests. Therefore choosing a pre-existing user model on the base of social data might lead to even bigger incoherence as discussed in the first case.

Kovacs et al have now chosen a really interesting way. They created a sample application for furniture namely chairs. Four users had to put 41 chairs in an order and evaluate their attributes, which gave them four user models. Based on those four models they are now able to generate for every new user a personal user model.

In the beginning every new user gets a set of 16 emotional words related to chairs. The user should now give his impressions about those emotions by selecting on a slider "+", "0" or "-". After the user’s adjustments the best fitting chair appears on the screen and after the user selected every of the 16 Kansei values the values of the chairs appear on the screen again with sliders to invite the user to make modifications according to his impressions.

In the background the system was calculating an approximate user model based on the four offline generated user models. This process is driven by fuzzy reasoning. The system selects the appropriate emotions and creates a fusion on the base of the pre-existing user models and returns a new user model based on the user emotions.

The only problem with this method is that the system will never learn something new. The system has always just the four pre-existing user models to create a fusion. This means to improve the system we’d need to create more basic user models.

3 Evaluation

3.1 Creating Accounts

To find the best possible user model for every new user is one of the most important but as we've seen in Kovacs as well one of the most difficult parts of a Kansei application. Many applications are not really focusing on this part and therefore get no sufficient results. Even K-Dime's way to find the right user model isn't really appropriate. It just assumes that every user with the same social background has the same idea for every, or at least most of the impression words. But there is no way to assure this assumption.

Even if we'd be able to find a way to create a user model there are still some disadvantages using K-Dime for an image search. Nowadays users are spoiled. They love applications where they can start with the first click and almost immediately get the result they want. But to initialize K-Dime takes a lot of time. Every user has his own vocabulary, words change during time and no application will ever be able to collect a complete set of impression words. Which means every user would have to spend a lot of time with the application without really using the real features of it.

On the other hand even if we'd succeed to find the best fitting user model with a complete set of impression words there is one more problem. As we've seen K-Dime creates a neuronal network by characterizing emotions or impression words. But users tend to use the same word in different situations. For example if we take the word "quietness" this could either be used as a peaceful feeling or in a different situation in a sense of loneliness. K-Dime would now adjust the dimensions of quietness every time it is used in a different way. This makes it quite impossible to create a coherent set of impression words.

3.2 copyright / privacy protection

With the Web 2.0 and all those new applications especially Facebook and Google a new aspect needs to be discussed. The point of privacy protection became more and more important in Medias agenda. Considering that an application like K-Dime is collecting a huge amount of personal data in a way we've never seen this might be one big reason why such an application would fail. In combination with the data collected by a social network or a mail service like Google we'd be able to interpret every single message in the exact way the user meant it as we know the users idea of every single impression word.

3.3 Performance / web application

Another problem with the K-Dime is its performance. In ME 79 they describe image processing as the most expensive operation. On thumbnails, image processing is relatively fast. On a Pentium 1.5Ghz, the filtering of 480 thumbnails takes 25 seconds". Our computers today have a lot more processing capability than 1.5 GHz, but should a user with an older computer try to process not thumbnails but whole pictures, it would certainly take very long. They state, that "The manual browsing of 40 pages, i.e. 482 images, on AltaVista takes about 10 minutes" (ibid.: 79), but they leave out two important factors here: On one hand it's more interesting to browse images for 10 Minutes than just waiting for a bar to reach 100% and on the other hand the resulting output of the K-Dime has to be browsed as well.

Concerning the results for the query "romantic, they state that "an average user had to browse up to 16 images before finding the first relevant image, up to 42 images for the second relevant image and up to 274 images to find a total of 8 relevant images" (ibid.: 78). They maintain that the average user has to browse only 20% of the images with K-Dime to find the same amount of relevant images: "first image a relevant one, while s/he has to browse 3 images to find a second relevant one and only 59 images to find a total of 8 images" (ibid.: 79). That is without question a good result for the K-Dime but I think, here they also don't answer a relevant question: How big is the convergence of the images a user would pick without the assistance of the K-Dime with the images found by the K-Dime. This is rather difficult to answer, but if they aim to support the users decision process (compare to MW: 103), it is a question that has to be answered.

The good performance of the K-Dime is not debatable with the provided data; there is still another small problem to be mentioned. At least for me, it's unnatural to use another software to filter the web, when I'm connected 24/7 and my browser is open at any time on the computer. And as even more devices are used to access the internet, a user would have to get the application on every device he uses. It would be better to be able to use the K-Dime over a web-based interface, without having to make a detour over the operating system. And this is where the high usage of computing power becomes a big handicap. Servers just don't have enough spare capabilities to do those calculations for every query a user makes.

And the competition doesn't sleep: If we take a look at http://www.google.ch/imghp, we can see that for searching images, Google has built in some of the features the K-Dime offers: Pictures can already be filtered through categories like "Face", "Photo", "Clip Art" or "Line Drawing" as well as through colour and resolution. It is not surprising, that Google creates a database with the so-called "low-level" features of an image to help users find "the right picture". Another take of Google on the issue of image searching is the "game" at http://images.google.com/imagelabeler/, where Google uses actual humans to process images and probably create more relevant search results.

Therefore I would certainly question, if the additional effort in creating and training a K-Agent can be a strong competition to the already existing ways of interacting with images on the internet (http://photosynth.net, http://www.flickr.com and http://www.deviantart.com to make three examples).

3.4 Conclusion

The idea promoted by the K-Dime, to tailor search engines for a single user to deliver an experience filled with the relevant information is definitely a very intriguing one. And as well we see the necessity to improve web searches within this huge flood of data. But because of the conclusion in the most current paper about the K-Dime we could find (MW: 105-106), and because Bianchi-Berthouze's more recent publications indicate a shift in her research topics, we assume that the K-Dime project has come to a conclusion for the time being. They were successful in proving that it's possible to increase the productivity in searching the web for images with the help of subjective categories. Furthermore they stated that users profit from the ability to externalize their own cognitive processes. And from the results they made with the K-Dime, they could also find some problems that future solutions for human-machine interaction will have to solve: Computers need to become able to model the variability of the user's state of mind as well as the strong dependence on their own context.

In this way they were also able to prove the usefulness of the Kansei model. But at the same time we have the Internet which is itself very dynamic and always in a process of restructuring itself [würdi nid säge, es duet sich nid restruktiere, das isch ja momentan grad das rise Problem im Web, dass es ebe immer no glich sturckturiert isch, wie vor 10 Jahr]. We have elements like the semantic web, rss/xml and content sharing through social networks which already point to a more personalized web.