Agents for Social Search in Long-Term Digital Preservation

Josep Lluis de la Rosa, Albert Trias, Raivo Ruusalepp, Kuldar Aas, Alex Moreno, Eloy Roura, Albert Bres, and Teresa Bosch

1EASY Innova & UdG, Parc Científic de Girona, Girona, Catalonia, EU
pepluis@eia.udg.edu (contacting author)

2National Archives of Estonia, Liivi 4, Tartu, 50409, Estonia
kuldar.aas@ra.ee

Abstract - This paper describes the application of agents to automate information exchange for digital preservation. Agents are able to recommend preservation solutions and also apply them to different preservation situations. Trust models for question-routing and answer ranking that are implemented by means of agents, show greater performance than traditional keyword search methodologies.

I. INTRODUCTION

Digital Preservation (DP) has become a ubiquitous problem that concerns everyone who has digital information that must be stored for over 15 years. While the use or prevalence of DP automation is low, the volume of digital information, diversity of digital formats, and types of digital objects is increasing.

Research in DP has moved away from trying to find one ideal solution to focus on defining practical solutions for different preservation situations [1]. These solutions must exploit the expert knowledge of memory institutions, be based on industry standards and, above all, be scalable and adaptable to disparate environments. Unfortunately, knowledge about DP is fragmented into disconnected islands of expertise. At this time, only large-memory institutions, such as libraries and archives with expert knowledge and specialized tools, have been tackling this problem by sharing the challenge with each other and with individual collectors and creators.

Agents are a fresh approach to overcoming the fragmented nature of expertise in DP. They are a design metaphor with anthropomorphic features for the implementation of software with proactive, social, and self-organizing properties.

II. AN AGENT-BASED SOLUTION

Collaboration to share knowledge regarding how and when to preserve digital objects already exists. Individuals and expert users (curators, preservation specialists, public, and private agencies) work openly to provide solutions, advice, and web services to enable long-term object preservation.

An open community sharing all types of resources (similar to wikis and blogs) applied to DP recipes for appraisal, transfer, file format conversion, and other DP methods is emerging.

and this community provides easier access to DP solutions.

DP is a collaborative activity where 67% of expert users look for solutions provided by other institutions, 90% of these users consult with trusted colleagues, and 83% consult with individual users [2]. As a consequence of its fragmented nature, there are frequent knowledge exchanges about DP, as 83% of expert users share their solutions with colleagues and individual users. Expert users feel that individual users are more familiar with practical DP needs thus they have the potential to widen the scope of DP solutions.

Social networks, Web 2.0, and their future automation are indications that information management is also becoming ubiquitous. One can observe an increase in freedom, specialization of work, and public availability of information through the Internet. This ever-growing universe of information requires proper management. Organizing this increase in information is no longer only a problem for organizations, but it is now an issue for individuals as well.

It is about everybody sharing information as an added-value service. Skilled DP experts should be available to share their knowledge with the vast majority of lay people who are often unaware of the need for DP, and social networking may be a way to bridge the DP knowledge gap.

Experts as a profession of information pathfinders [3] are becoming relevant today for DP. Blogs and Web sites provide a reference function by connecting people each other and with other resources, and libraries are recognizing the value of trying to assimilate “social networking” into their reference systems. These changes in the reference world will produce a permanent flow of innovation and will invite new information agents from outside the current network of DP professionals.

This change is a promising and enriching trend, especially for our approach if intelligent agents are managing the flow of information and learn to manage DP.

The use of social networks for DP seems not only plausible but even necessary. Because of its novelty, DP is receptive to many personal experiences and opinions. We have termed these experiences and opinions DP “recipes”. Gosain and Liljenback [8] [12] have studied how people are encouraged to exchange solutions within communities through means such as company intranets or question-answering sites (QA).

DP recipes will connect users with similar needs and who might benefit from the same recipes or better understand or apply those recipes. Today’s strategies for DP are labor-

1 http://ercim-news.ercim.eu/en80/keynote January 2010
2 Recipes are preservation action plans that are developed with the particular experience of a user that potentially a number of users will find that this solution works for them as well.
intensive and often require specialized skills, especially when dealing with complex digital objects. Users are best equipped to understand the needs of other users, and therefore the goal is to match people effectively and let them share. A network of people providing each other multiple DP solutions has been described as crowdsourcing [9] or social intelligence which is defined as human participation in the gathering, synthesis, and dissemination of knowledge and problem solving.

Thus, more and more people are becoming active users contributing to social intelligence that solve problems by crowdsourcing. One of the most important challenges of the next decade will be to determine how to design systems that will enable this sort of socially intelligent collective activity, as for example [5] [15] for crowdsourcing question-answering. Agents are a good approach to facilitate such design systems, as agents are actor centered and can emulate human behavior and social skills by gathering, synthesizing and disseminating search hints that are useful for social search. The bottom line is that personal agents will provide knowledge for preserving digital objects.

Precedents of multi-agent architectures exist that deliver more relevant and better personalized answers to the user than existing QA search engines. These modern architectures are a means of efficient information exchange between humans [7]. For example, peer-to-peer search recommendation is a synergistic approach, and the prototype Sixsearch.org 2.0 [11] takes into account not only which neighbors are the best resource providers for a given query but also which combinations of neighbors can provide the least redundant results [19] in question routing.

III. WHY SOCIAL SEARCH WITH AGENTS IS MORE EFFECTIVE THAN EXISTING SEARCH ENGINES FOR DP

Adding context to user queries by means of semantic web, personalization, or deeper grammatical comprehension improves the relevance of the answers\(^5\). However, users are reluctant to trust and reuse answers from search engines unless they know that similar users or colleagues have applied the answers successfully. This social feature is an integral part of the agent metaphor [13] used for DP.

Google is an example of an existing search engine that is useful for finding answers about DP by indexing Question-Answer (QA) sites as for example Yahoo! Answers. However, finding relevant answers to DP questions using existing search engines is still a problem. While Google remains the dominant search engine for general QA searches, common on-line users are migrating to online communities such as those found at MySpace, Craigslist, Facebook, Twitter, Orkut, Myweb and ezboard. Each of these emerging community networks provides search options [16], and users searching for DP information may cluster around DP communities in the future. With Facebook now receiving nearly twice the daily traffic as Google, search engines are no longer the exclusive internet search tool for users. In emerging online communities, users need to establish accounts and set profiles, which requires essential skills such as using word association and applying knowledge relevant to local searches. To make information available to various search engines, users are also required to learn about tagging, linking, and blogging so that spiders and bots will be able to find the information.

A PROTA GE report [2] said that 77% of expert users search the web for DP solutions. In addition, 60% of expert users visit hundreds of web sites relevant to DP, to whom they trust, namely the National Archive, ISO, Skydrve.msn.se, PTC.com, sun.com, Dassault, DPC, JISC Digital Media, Planets Project, Microsoft, Autodesk, Project Gutenberg, and pdfa.org. Twenty percent of expert users contribute to DP web sites, which may seem like a low percentage. However, this is a high percentage compared with that of Wikipedia (which has less than 1% of contributions). This statistic confirms that the DP community is willing to share, and this encourages having a platform for expert users to share DP solutions, that would be used by institutions and individual users.

In social search, a chain of trusted agents matches a question from one user with a proper answer from another user. This type of searching will likely outperform the Google-like searches. Both users can be represented by their agents in favor of higher levels of automation in the process of matching questions. Every agent in the chain adds contextual and intentional comprehension of the question, and then chooses the next agent to find the answer by following its own trusted contact list. Agents will motivate their users to supply them with knowledge and referrals to other people and agents. This paradigm is an emergent approach for collaborative searches, and preliminary studies have been published by Freyne et al. [6]. Their works show how Collaborative Filtering (CF) methods exploit a graded mapping between users and items, and the I-SPY algorithm utilizes a similar relationship between queries and result pages (Web pages, images, audio files, video files, etc.).

A. Proof of concept: CF search (Google-like) performs worse than social (agent-like) search

A CF search algorithm discovers patterns in the activity of a community of searchers, uses these patterns to determine the general search context, and prioritizes the search results [6]. It makes no strong assumptions about the form of the underlying search engines, and the algorithm is applicable across a range of content types. The CF ranking metric requires no additional parsing of the result pages. The ability to personalize search results for the requirements of a community is achieved without the need to store individualized search histories. No individual user profiles are stored, and no user identification is necessary. If the user is identified, then the ranking increases in relevance. Google creates a user profile based on the users’ questions, and then compares it to other profiles so that the ranking of answers is influenced by similar profiles. This mechanism is comparable to CF of searches where one user receives the same answers (A) as others who had similar questions (Q).

\(^5\) See one example compared to the Yahoo search engine http://box.cs.rpi.edu/wise/compare.jsp
Some users’ answers may be more authoritative than others, and this is obviously important in the context of CF searches. Trust incorporates varying levels of user expertise, and this is one of the social features of agency [13]. A trust-guided search uses contact lists and explores trusted profiles. Thus, rather than requiring a single similarity function F that needs tuning, a trust-guided search uses the contact lists to conform to several similarity functions F’, which are created from unique, personal points of view. The combination of personal points of view produces better results than a single F.

Movilens4 is used for a proof of concept [4] [14]. It is a source of ratings as a special case of question-answering to which a trust-guided search can be applied. We used 100,000 ratings from 943 Movilens users. 80% of the users were for training, and the remaining 20% were for testing.

In the experiment shown in Table I, each user had an agent. Using the training set, an agent asked for a rating on a subject (movie) at every simulation step. The CF that chose the most similar users was then compared to a trust-guided recommendation that selected the answer from the most trusted user in a contact list size of 2. A random recommendation was the baseline value.

<table>
<thead>
<tr>
<th>TABLE I. COMPARISON OF TRUST SEARCH VS. CF SEARCH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Trust</td>
</tr>
<tr>
<td>CF</td>
</tr>
<tr>
<td>Random</td>
</tr>
</tbody>
</table>

The trust-guided algorithm resulted in an MSE value of 1.03, which was better than the MSE of CF (1.33) and the MSE of the baseline (2.09). The Trust-guided MSE was 22.4% smaller than that of the CF. We confirmed that our results were relevant because the student’s T-test results were significantly smaller than .05. The simplest trust-guided implementation selected the best answer in a social network.

This result is consistent with F. Martin’s recommendation #7, which was given at the 2009 RecSys conference in New York3 in the talk, “Don’t waste time calculating nearest neighbours if you have a social network”3 (you have it as long as you have the contact lists of users with their followers and friends on Twitter, Facebook, social networks fan sites, etc.). Contact lists, especially if ranked by trust, are a highly relevant primary source of recommendations.

IV. TECHNOLOGY DESCRIPTION – TRUST MODEL FOR DP

When there are several experts (and agents that answer on behalf of them) in several DP areas, it is necessary to know the relevance and trustworthiness of the information. The trust that an agent has in others depends on past interactions, and trust helps select who to ask for action plans or “recipes”. The agent can also use the existing reputation of others as a means of rating agencies. The core terms used in the rating model are as follows:

- Rated entities: services, institutions, companies, users or products that could be described on the Web
- Rating criteria: the pre-defined criteria used by an agency to rate an entity.
- Rating certificate: certificate issued to a rated entity to ensure its rating.

Ratings are independent third parties that are able to objectively evaluate DP services, institutions, companies, users or products based on standardized criteria. These agencies issue certificates to guarantee the authenticity of ratings. Ratings provide additional information for decision making (e.g., whether to accept or reject a “recipe”). Users decide which agency to trust and use, and they register on a voluntary basis with several rating agencies. The certificate is recognized by a partner only if the partner trusts the agency. The value of the certificate is also related to the reputation and the circle of trust of the rating agency issuing it. Outside the social network of the agency, the certificate is not recognized. To enable inter-community trust and collaboration, rating agencies are conceived to issue certificates of trust within certificates issued by other rating agencies.

Asknext [18] is an agent environment with a novel question routing method. It implements the described trust model by means of ranked contact lists. The inaugural members of the contact list are rating agencies that are kept on the contact lists while their recommendations stand above other agents.

V. FUNCTIONALITY OF AN AGENTS PLATFORM prototype

In the prototype there are three users (Eloy, Albert, and Alex) and an Access Point with certified DP actionplans. Their contact lists are the following:

- Alex’s: {Eloy},
- Eloy’s: {Albert and Alex},
- Albert’s: {Eloy}

They have several actionplans in their collections (Table II).

<table>
<thead>
<tr>
<th>TABLE II. THE ACTION PLANS OF THE THREE USERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ActionPlan</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Image conversion (3versions)</td>
</tr>
<tr>
<td>Tech metadata extraction</td>
</tr>
<tr>
<td>Remote AV Check</td>
</tr>
<tr>
<td>Remote AV Clean</td>
</tr>
<tr>
<td>Local AV Check</td>
</tr>
<tr>
<td>Local AV Clean</td>
</tr>
<tr>
<td>Remote AV Check</td>
</tr>
<tr>
<td>Remote AV Clean</td>
</tr>
<tr>
<td>Generic Image Conversion</td>
</tr>
<tr>
<td>Calc MD5</td>
</tr>
</tbody>
</table>

Alex needs an actionplan. He types the keyword “AV” to check for viruses in a picture of his collection called “girona_at_night.jpg”. The process including the social search by the agents is as follows:

1. Alex searches the actionplan keyword in the “Application local DB” (degrees of separation = 0). No result is found.

3 http://www.movielens.org
4. Alex’s Agent searches for a certified actionplan into the Access Point (institutional point of view, with certified plans). No result is found again.

3. The Agent uses ASKNEXT to search the social network. It asks its friend Eloy (degrees of separation = 1).

4. Eloy’s agent checks whether 4 matching plans are good for Alex (trust-guided decision). Two plans (collection and author) do not match; their fulfillment \( F = .35 \) is lower than the QoS of .50 (the Quality of Service threshold).

On the other hand, two trusted plans match (fulfillment \( F=1.00 \) as author and collection match higher than .50 of the QoS). Eloy’s Agent sends these 2 actionplans to Alex.

5. Alex’s Agent receives the actionplans from Eloy’s agent and ranks them for Alex.

6. The plans are added to Alex’s actionplan collection and are tagged “Eloy” as provider and author. According to the actionplan profile configuration, QoS, and the maximum degree of separation (3), the agents were able to recommend the Local AV Check and Local AV Clean actionplans to Alex. Alex will then evaluate the received actionplans and update his trust in Eloy.

VI. CONCLUSIONS AND RECOMMENDATIONS

We have outlined how the agent metaphor helps the exchange of ongoing knowledge and solutions for DP. Agents can evaluate DP solutions and proactively execute them.

Finding relevant DP answers is difficult when using standard search engines, and adding context to user queries by semantic web, personalization, deeper grammatical comprehension or trust improves the relevance of answers.

In social search, chains of trusted answers of trusted agents match a question from one user with a proper answer from another user. Both users can be represented by their agents resulting in higher levels of automation in the process of matching needs. Every agent in the chain chooses the next agent to find the answer by following its own trusted contact list. Agents encourage their users to provide them with knowledge and referrals to other people and agents. Our proof of concept showed that the trust-guided search outperformed the CF approaches by finding the most suitable answer."

Agents provide an advantage over existing search techniques for many reasons. First, direct QA matching has a lower relevance than a chain of intermediate contexts by agents. Second, agents reach the Deep Web. Third, Google-like searches explore (crawl) a flat search space, while agents explore an intentional search space. Finally, every user provides knowledge with the user’s own ontology, and agents naturally do the same while current search engines cannot.

Our prototype showed that users can manage their DP knowledge exchanges by means of agents answering on their behalf, which is definitely a workable approach.

The **limitation of our approach** lays in the evaluation of answers because one cannot know in advance the long-term results of a DP recipe. Delayed feedback can be simulated and experimented in two ways: first, as erroneous feedback in a certain percentage of users; and second, by introducing a delay into the feedback. The little research on delayed trust is Leake [10], Klos [17], and Bos [20], that will be applied to solve the lacks of our trust guided approach applied to DP.

ACKNOWLEDGEMENTS

This research was partly funded by the EU projects Num. 216746 PREservation Organizations using Tools in AGEnt Environments (PROTAGE), and Num. 238887, a unique European citizens’ attention service (iSAC6+), and the UDG research grant BR09/10.

REFERENCES


