Assessment of a Proprietary Online Smart-Family-Matching Tool to Reunite Lost Families.

Juan Rojas
School of Electrical and
Computer Engineering
Sun Yat-Sen University
Guangzhou, Guangdong 510006
Email: juanlvis@mail.sysu.edu.cn

***Abstract*—The presence of wars, civil strife, and natural disasters inevitably separate many families from their loved ones. Existing tools to help divided families find loved ones consist of basic meta-data searches that are often unable to return a match in the face of incomplete or inaccurate information. This paper presents a first-order analysis of a proprietary online system that used correlated meta-data, genealogical information, and single images to search and match lost relatives more effectively. Experimental results showed that the system reliably guaranteed the right match in the presence of complete and accurate information. The results also showed that when incomplete meta-data information is present along with genealogical information, false-positive matches occur within one's own family tree. Furthermore, with increased genealogical data, the likelihood of false-positives can decrease by almost 44%. Image processing tools did not improve the results of the match when only one image is used. A larger database of family pictures could improve the likelihood of finding the right match. This paper assessed the viability of the system as a tool for humanitarian organization that help refugees located loved ones. Based on the results, we recommend the integration of genealogical data into existing search systems used by humanitarian organizations.**

I. Introduction

Every year families are divided due to war, civil strife, and natural disasters. Family members, in the face of natural disasters, are often forced to relocate to shelters; or in the face of armed conflict to relocate to camps [1]. The United Nations High Commission for Refugees (UNHCR) reports that in 2010 there were an estimated 25 million people forcibly displaced. From these, 15 million were internally displaced people (IDP's) and 10 million were refugees. Of the refugee population, 80% lives in developing nations– primarily in Africa and the Middle East [2]. Half of the refugee population also lives in urban areas. In 2011 alone the UNHCR will budget $3.3 billion USD to support forcibly displaced populations [3]. Moreover, in the recent crisis in Japan, more than 17,000 people are estimated to be lost. Not only do victims of disaster deal with the loss of possessions and their livelihood, but more significantly they struggle to find their loved ones.

According to the UNHCR's statistical yearbook [4], local governments are responsible to register displaced people. If they are unable to do so, the UNHCR can assist. The UNHCR implemented ProGres, a registration software that is now used in more than 75 countries to keep better records of displaced individuals [5]. While ProGres and similar databases for registering refugees have developed significantly, it is unclear how useful they are to find lost relatives [6]. Apparently, the number of tools available to find missing family members varies with income levels in the region. For example, in the aftermath of the Japan tsunami in March 2011, local governments and organizations have compiled many technology-rich resources to find missing loved ones, including online databases and call-lines [7], [8]. RefUnite, is example of an organization trying to use technology to make the search easier for divided families. The organization recently made available an online search tool for registered refugees [9]. All surveyed databases allow people to include personal information including: names, addresses, contact info, and locations where last seen. The effectiveness of these search databases, however, depend on the accuracy of registered data on both sides of search process. As far as the author understands, the search databases perform a one-to-one search. That is unless there is a match between any two field entries, the search will return no results. None of the surveyed searches involve features like comparing genealogical data, facial features recognition, or variations in spellings of names. Besides registrants meta-data, considering genealogical patterns, facial feature similarities, and spelling variations could significantly enhance the search process effectiveness.

This work assessed the effectiveness of a proprietary online family-search tool with smart matching capabilities. The software is hosted by the "www.myHeritage.com" website [10] and runs a search algorithm that matches members based on meta-data, genealogical tree relationships, and facial similarities.

The rest of the paper is organized as follows. Section II describes the details of the proprietary software. Section III describes experiments and results carried out to assess the matching capability of the proprietary tool. Section IV

discusses various considerations of using such a software. Section V summarizes key findings and suggests directions for future work.

II. Description of Proprietary Software

The online proprietary software that is studied in this paper was developed by an Israeli based company ("www.myHeritage.com"). The website is designed to build family trees by having individual family members include their personal and genealogical data, along with family member images. The creators of the site believe that as their website grows, distant family members, who may not have contact with each other will register under their website and find each other through their smart search software.

A. Registration

Registration on the site requires completing a few fields of personal information. On the home page, one can register by including personal information consisting of gender, first name, last name, email address, country of origin, and the last name of the father. Optional information includes personal year of birth, father's first name, and mother's first and last names. After filling the appropriate information, one clicks "Go" and is taken to a choice between a free plan or a premium plan. This study was effected under the free plan. The third step includes selecting and confirming a password. After that a "Family Site" appears.

B. Family Tree

For the purposes of this study, we are interested in two sections of the website. Both sections of the website can be found on the 2nd menu tab on the top-left corner of the site under the name of Family Tree. The sections of interest are: (i) "Tree" and (ii) "Smart Matches". Fig 1, shows an example of a tree in the system.

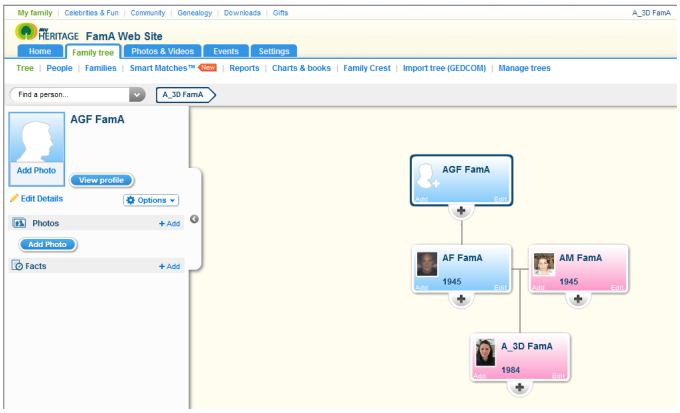

Fig. 1. A depiction of a family tree on the "myHeritage.com" website.

1) Tree: The "Tree" section of the site includes a user friendly interface to develop a genealogical tree. The standard initial setup includes at minimum two boxes indicating the relationship between the registered participant and the father. If the mother was included in the registration portion of the site, a box representing that familial connection also appears. Each box can show up to 5 pieces of data per family member: (a) Gender: males are shown in blue colored boxes, females are show in pink colored boxes; (b) first name; (c) last name; (d) year of birth, and (e) a picture.

Each box in the genealogical tree can be edited by clicking one of three links. Two links are positioned in the lower left and right corners and are named "Add" and "Edit" respectively. A third link has the form of a plus-sign and is located underneath each box in the central part of the latter. Editing capabilities allow to change personal details; add or remove familial relationships in tree (including parents, partner, siblings, or children); and managing pictures.

This paper uses a mathematical representation to denote family trees, members in the family trees, and the personal identifying information of each member. Family trees are represented by the set π_{i_j}, where i corresponds to the family member that created the tree, and j corresponds to the family the tree represents. For this study we limit the scope of a family unit to size four. The family members along with their abbreviation are: father (**f**), mother (**m**), son (**s**), and daughter (**d**). Each family tree is composed of family members ϕ_k, where k denotes the representative family member. Each family member in the tree is characterized by five personal data items: *First Name* (**FN**), *Last Name* (**LN**), *Gender* (**G**), *Birth Date* (**BD**), and *Picture* (**P**).

Using encoded information, a family member's available data set(it does not have to be the complete set) in a tree can be included as:

$$\phi_k(FN, LN, G, BD, P). \tag{1}$$

Furthermore, a family tree created by the father of a given family, A, that includes full information about his wife (the mother), son, and daughter is represented as:

$$\begin{aligned}\pi_{f\_FamA} = &\phi_f(FN, LN, BD, G, P) \\ &+ \phi_m(FN, LN, BD, G, P) \\ &+ \phi_s(FN, LN, BD, G, P) \\ &+ \phi_d(FN, LN, BD, G, P).\end{aligned} \tag{2}$$

2) Smart Matches: The "Smart Matches" section of the website is the second area of interest [11]. The website does not disclose the search algorithm in detail, but it states that it runs on two fundamental technologies:

- Smart Matching: Finds similar profiles across family trees stored in their database [11] based on meta-data information and genealogical information. Meta-data is compared to see if different profiles or family members in trees share the same first and last names, date's of birth, gender, and country of origin. Genealogical data is

compared to see if the number of parents, siblings, and children matches across profiles.

- Face Recognition: Compares facial images of registered members with all faces previously stored in their database. The facial search is enhanced with an increased number of pictures of the applicant and with manual annotation of the name of the applicant. The software claims to work even at different ages of the same person's life.

The proprietary software may offer matches even if all the information does not agree. The matching software also provides a "rated match likelihood estimation". The exact computation of the likelihood estimation is not described but it compares at least twelve factors that are disclosed as: gender of registrant, first and last names of registrant, birth and death date (if available), the father's and the mother's first and last names; and any spouses, children, or siblings if available.

Experiments test whether the proprietary system can find matches of family members across two or more independently created family trees. In this paper, we restrict searches to include two family trees. In two family trees, there can be one-way searches or two-way searches. One-way searches refer to searches only initiated by one party. Two way searches refer to searches initiated by both parties. Consider a family "A", where the mother generated family tree, π_{m_FamA}, that included full information about herself and her daughter, then:

$$\pi_{m\_FamA} = \phi_m(FN, LN, BD, G, P) + \phi_d(FN, LN, BD, G, P). \quad (3)$$

Similarly, if the daughter generated a tree, π_{d_FamA}, with full information about herself and her mom, it could be represented by Eqtn. 3. This case scenario is depicted in Fig. 2, where the mother's and daughter's tree are shown on the left and the right respectively.

$$\pi_{d\_FamA} = \phi_d(FN, LN, BD, G, P) + \phi_m(FN, LN, BD, G, P). \quad (4)$$

Then match criteria is a function of the relevant family trees and the person(s) to match. In a one-way match, one must specify relative to what family tree the comparison is being made. In the above example, if the mother searches for the daughter, the meaningful match would be relative to the mother's tree, not the daughter's own tree. In the same way, if the search is for the mother, the meaningful match would e relative to the daughter's tree. An one-way match function, $m(...)$, includes a subsidiary family tree, π_s, which is compared against a base tree, π_b, for the person, ϕ_1, and produces a likelihood percentage, c, as in Eqtn. 5.

$$m(\pi_b, \pi_s, \phi_1) = c, \quad (5)$$

In the case of the mother and daughter, the match function would look as:

$$m(\pi_{m\_FamA}, \pi_{d\_FamA}, \phi_d) = c. \quad (6)$$

For two-way matches, the function would include two people: the first person, ϕ_1, would be the match of the subsidiary tree relative to the base tree, and the second person, ϕ_2, would be the match of the base tree relative to the subsidiary tree. The match function would also produce two likelihood percentages: c1 relative to the first comparison, and c2 relative to the second comparison. The two-way match function is expressed as:

$$m(\pi_{m\_FamA}, \pi_{d\_FamA}, \phi_1, \phi_2) =< c1, c2 > . \quad (7)$$

III. Experiments

Four experiments were designed to assess the basic performance of the smart search's fundamental technologies: genealogical data and face recognition with incomplete data records. The conducted experiments focused on representative case scenarios that can showcase the system's performance in the presence of incomplete family records. For these experiments, pictures from a volunteer family consisting of a father, mother, son, and daughter were collected along with relevant personal and contact information. A separate profile was created for each member of the family.

The first experiment served as a baseline experiment for incomplete data records. The family trees of each member, include only one family member, namely oneself. Additionally, the personal data record is incomplete. The second experiment tested how the estimation likelihood of finding a right match changed as more family links were inserted into the tree of two family members. The third experiment studied the response of the system in the case pictures of family members were included as part of personal data record. The last experiment, served as a baseline against complete and accurate data records.

A. Experiment 1

In experiment 1, four profiles were created: $\phi_f, \phi_m, \phi_s, \phi_d$, each of which did not include pictures. One piece of data was omitted at random from each of the members: $\phi_f(FN, LN, G)$, $\phi_m(LN, G, BD)$, $\phi_s(FN, G, BD)$, and $\phi_d(FN, LN, BD)$. After the profiles were created, the family tree was configured to only have that one member in their own tree.

B. Experiment 2

In experiment 2, focus was placed on smart matches with extended family trees of only two family members: the father and the son. The family trees π_f and π_s were expanded to include one more family member in an incremental manner. With each added member, we ensured that the latter be present in both trees and with the full and accurate personal information on the other. For a first order study, random selection of members was not tested, nor was other combinations of incomplete personal data. In effect, the first addition consisted of adding the son's profile to the father's tree and the father's profile to the son's tree. Afterwards, the mother was added to both trees, and finally the daughter.

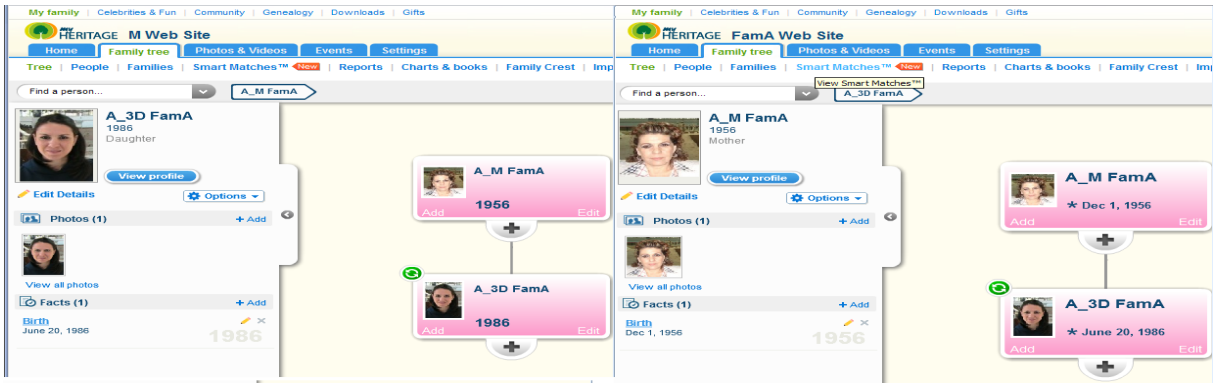

Fig. 2. A side-by-side comparison of two family trees including a mother generated tree on the left and a daughter generated tree on the right.

1) First Case: In the first case, the father's family tree was composed according to Eqtn. 8 and the son's tree was composed according to Eqtn. 9.

$$\pi_{f\_FamA} = \phi_f(FN, LN, G) + \phi_s(FN, G, BD), \text{ and} \quad (8)$$

$$\pi_{s\_FamA} = \phi_s(FN, G, BD) + \phi_f(FN, LN, G). \quad (9)$$

2) Second Case: The second instance of the experiment, adds the mother on both trees:

$$\pi_{f\_FamA} = \phi_f(FN, LN, G) + \phi_m(LN, G, BD) + \phi_s(FN, G, BD) \text{ and,} \quad (10)$$

$$\pi_{s\_FamA} = \phi_s(FN, G, BD) + \phi_f(FN, LN, G) + \phi_m(LN, G, BD). \quad (11)$$

3) Third Case: The third instance adds the daughter:

$$\pi_{f\_FamA} = \phi_f(FN, LN, G) + \phi_m(LN, G, BD) + \phi_s(FN, G, BD) + \phi_d(FN, LN, BD), \text{ and} \quad (12)$$

$$\pi_{s\_FamA} = \phi_s(FN, G, BD) + \phi_f(FN, LN, G) + \phi_m(LN, G, BD) + \phi_d(FN, LN, BD). \quad (13)$$

The two-way match is trying to match the son in the father's tree and father in the son's tree:

$$m(\pi_{f\_FamA}, \pi_{s\_FamA}, \phi_s, \phi_f) =< c1, c2 > . \quad (14)$$

C. Experiment 3

The third experiment follows the same steps of experiment 2. The difference is that in this occasion a picture was included for all family members. The same picture was added for members across different family trees.

D. Experiment 4

The last experiment follows the same sequence as the previous two experiments for the same two family trees. In this instance, we assume that both family members registered their families with complete and accurate information. Namely the set (G, FN, LN, BD, P). This last experiment was expected to give much higher likelihood estimates that any of the other tests and could serve as a basis for accurate results.

Fig. 3. Screenshot of a false-positive match in the father's family tree with himself.

IV. Results and Discussion

This section presents the results for four executed experiments.

A. Experiment 1

The system returned a smart match for each of the four profiles, albeit with different likelihoods and an unexpected result.

For the father's profile, there was a single match, suggesting the match was itself but pointing to the son's tree. In other words, the system purported a match function of the type: $m(\pi_f, \pi_s, \phi_s) = -43\%$, where the negative sign represents the false-positive. The system matched the gender in the other tree and must have found some correlation in the names. The system presented mismatches for the first names of both profiles as well as the birth dates and no available data for the other fields. While the match was incorrect, the likelihood was under 50%. A screenshot of the match in the father's tree is shown in Fig. 3. Similarly, in the son's profile, there was an incorrect match, suggesting the match was with itself but pointing to father's tree: $m(\pi_s, \pi_f, \phi_f) = -41\%$. The same explanations were presented by the system, it is unclear why the system presents a variation in the match estimation likelihood.

For the mother and for the daughter's profiles, there were matches, but these pointed to the family tree belonging to the daughter and mother respectively. In effect, it returned the name of the matched family tree and the person that created it, but there was not a direct correlation with a person. This was unexpected.

B. Experiment 2

For the second experiment, genealogical information was used for the first time, albeit with incomplete personal data.

1) First Case: In the first case a match was recorded in the father's tree and the son's tree. Both were false-positive matches, suggesting that the system believes to have found the same person in the other family member's tree. Here the system returns a two-way match: $m(\pi_f, \pi_s, \phi_s, \phi_f) = < -36\%, -31\% >$.

The likelihood estimates are lower in both cases indicating an improvement by the system to tell that these are less likely matches. The improved estimation came from the familial relationships in the trees. In the father's tree, the father has a son, while in the other tree the matched person has none. The same idea followed in the son's tree.

It is also useful to highlight that the system returned another piece of information which states that there was a match between family trees, with two family members in each one.

2) Second Case: In the second case the same matches were recorded as in the first case. There were false-positives in both cases. The likelihood estimate remained the same for the father's tree but diminished for the son's tree from -36% to -31%. The system recognized that in the father's tree, the father had a wife but not the match in the other tree. This difference, however, did lower the likelihood estimation in the son's tree. It would seem to the authors that the likelihood estimation for the father's tree should be lower in this case. What that is not the case remains uncertain.

As in the first case, the system stated that there was a match between family trees, but in this case it noted that each three had three members.

3) Third Case: For the last case, the false positives remain in both trees, but once again with lower likelihoods. The estimation likelihoods in both trees recorded a low percentage of -26%. The system recognizes that in the father's tree, the father has 2 children but none in the other, and that the father has no siblings while the member in the son's tree does have one. The opposite logic follows for the son's tree. The system also noted that both of these family trees contained four members each. We can conclude that while the right match was not yet ascertained, the system did reveal with each addition of a family member to the tree, that the likelihood of a right match was lower.

C. Experiment 3

Surprisingly enough, all the results for the third experiment yielded the same false-positives and likelihood estimates as in Exp. 2, for the three cases. That is, for case 1, where there is only a father and son in the tree, the likelihood that the father is the son in the other tree is equal to -36%. In case 2, where the mom was added to the family tree, the system estimates the false positive at 31%. Finally, the last case yielded a -26% of likelihood that the match was write as before. It appears as if the image processing task requires more time or more images or both to train from (website support staff would not provide detailed information).

D. Experiment 4

The last experiment, as expected, returned correct matches with high estimation likelihood values. In the first instance, where the father's tree and son's tree have complete and accurate information of each other, the two-way match function returned the following percentages: $m(\pi_f, \pi_s, \phi_f, \phi_s) = < 95\%, 85\% >$. The system correctly matched the son's in the father's tree, and the father's in the son's tree. The estimated likelihood was very high. Missing genealogical data included missing mother's and spouses. The system also matched the family trees themselves suggesting they are the same family containing two members each.

In the second case, the first 100% likelihood estimate is computed: $m(\pi_f, \pi_s, \phi_f, \phi_s) =< -64\%, 100\% >$. The system believes that the children in both trees are the same person. All 12 markers of personal and genealogical data match supported by similar facial features. The results concerning the father was unexpected. The system states that the mother's last name in both trees remains private to the other family tree. The reason is due to the distinction between the mother's maiden name and married surname. The system is looking for a clear match between those to names to distinguish the relationship between mother-son and wife-husband relationships. This ambiguity leads the system to provide a likelihood estimation suggesting that the son in the son's family tree could be a match for the father in the father's family tree and vice-versa. The system also matched the family trees themselves suggesting they are the same family containing three members each.

In the last case, the perfect match for the son occurs again and false-positive likelihood estimate for the father is reduced from -64% to -59% : $m(\pi_f, \pi_s, \phi_f, \phi_s) =< -59\%, 100\% >$. The system improves its prediction with the added information from the sibling/daughter relationship, but it cannot make the right match due to ambiguity in the mother/wife relationship. The system again matched the family trees themselves suggesting they are the same family containing four members each.

E. Discussion

The proprietary smart-matching tool of "myHeritage.com" seems a promising tool. This paper contributes a characterization of the basic advantages offered by this genealogical and facial feature search tool. In our discussion an evaluation of the performance of the system along with its advantages is presented. A discussion on its viability is also presented.

With respect to its performance, the system was unable to correctly match the two correct family members across family

trees but this seems to have been caused by the absence of the last name for the son. Nonetheless, the system was able to identify the correct family tree from the database. In all cases, it matched the father with the son or vice-versa for the second family tree. The experimental results provide evidence that the system learns better that the proposed match is not the correct one as more genealogical information is available. This is an important capability, as it is absent from all the systems used by organizations to find lost family members. With respect to image processing advantages of facial features, there were no advantage seen under our experiments. There may be two leading reasons: (i) the system requires more time to analyze facial features across the website's database (as of April of 2011 about 17,000,000 family trees entrees with an undisclosed number of pictures), and (ii) only one picture was uploaded per family member. In general, the genealogical information helped the system increase it's likelihood estimation almost 44% (from 46% to 26%). The system also showed that when complete and accurate personal information is provided, the number of false-positives decreases and the chances of getting a perfect match also increase as more genealogical information is made available. The experiments also showed how the mother member of a family may be more prone to ambiguity given that there is a mother's maiden name and a mother's married surname, both of which are unique and distinct and tell about the relationship that members holds with respect to siblings and the husband within the trees.

One useful tool that was not discussed during the experimental section, is a manual family tree comparison feature, which was available when there was a match across family trees. This tool, while not automated, allows the users to manually check if another person in the matched family tree matches for the searched relative. This option can prove very useful in cases where incomplete or inaccurate data provide false-positives but the matches are within the family unit of the searched family member. Testing was not conducted to see if the system would then learn by providing the right match.

The author deems it useful to further characterize the system via an exhaustive and comprehensive set of alternatives for the search of family members. That would include all combinations of incomplete or inaccurate personal family data, in combination of varying genealogical ties in family trees, and the absence and presence of one or more images.

The system as it stands provides advantages compared to systems used in actuality. The system is smarter than one-to-one searches; it also allows family members to input data as they deem fit. That is, if for security reasons two family members chose to use their nicknames instead of their official names, the system would behave in the same way.

The viability of the system would depend on a fairly unrestricted access to the internet in communities where the search tool is needed. While this kind of infrastructure is unavailable in most rural camps, there are non-profit organizations like the Jesuit Refugee Service that are working to change that fact [12]. Additionally, as stated in I, refugee data suggests that about half of the world refugees live in urban areas. Kampala, Uganda is a city where many of these urban refugees live. In a city as Kampala, access to the internet is readily available through the many internet cafe's that sprawl the city. Furthermore, in cases of natural disaster in developing or first world nations, sheltered communities can more easily access the online resource to search for loved ones (the fact that the website can be used in a host of languages also helps).

In effect, however, if such a system were to be used in large volumes it would have to be standardized as part of the search tools used by the largest and most prominent humanitarian organizations. The UNHCR would do well to integrate such a system to their ProGress refugee registration system, and so would Google.org for its people finder system. Other groups like RefUnite would also provide a richer search medium to their users.

V. Conclusion

This paper presented an assessment of an online proprietary tool used to search lost family members that not only uses personal contact data but also collects genealogical data and images from participants. The assessment found that based on meta-data alone, the system is able to match candidates to at least the same family tree. Incremental amounts of genealogical data help the system increase the accuracy of its likelihood estimation nearly 44%. The image processing aspect of the search appears to need more time and training images to make a difference. Recommendations for more comprehensive characterization of the proprietary system and standardization into the already existing tools of humanitarian organizations are presented.

VI. Acknowledgment

The authors would like to thank the site "www.myHeritage.com" for the ability to use their basic program free of charge.

References

[1] C. ODwyer, "Family reunification for refugees and the application of section 29 of the family law act 1995," *The Researcher*, vol. 6, no. 1, March 2011. [Online]. Available: "http://www.unhcr.org/refworld/docid/4d8b39282.html"
[2] UNHCR, "Global trends," UNHCR, Tech. Rep., 2009.
[3] ——, "Global appeal 2011 update," UNHCR, Tech. Rep., 2011.
[4] "Unhcr's 2009 statistical yearbook," UNHCR, Tech. Rep., Oct. 2010.
[5] *ProGres, UNHCR's Registration Software*, April 2011. [Online]. Available: "http://www.unhcr.org/cgi-bin/texis/vtx/search?page=search&docid=4c342fde6&query=proGres#hit1"
[6] *ProGres*, July 2010. [Online]. Available: "http://www.unhcr.org/4c342fde6.html"
[7] *Google People Finder*, April 2011. [Online]. Available: "http://www.google.com/intl/en/crisisresponse/japanquake2011.html#resources"
[8] *Japan People Finder*, April 2011. [Online]. Available: "http://sosjapan.org"
[9] *Refugee Search Tool*, April 2011. [Online]. Available: "http://www.refunite.org"
[10] *myHeritage Website*, April 2011. [Online]. Available: "http://www.myHeritage.com"
[11] *MyHeritage-SmartMatches*, April 2011. [Online]. Available: "http://www.myheritage.com/FP/smart-matching.php?s=147454131"
[12] M. Macchiavello, "Livelihoods strategies of urban refugees in kampala," Tech. Rep., 2004.