Ethical intelligence in social recommender systems

Peter Juel Henrichsen
Center for Computational Modelling of Language
Copenhagen Business School
pjh.isv@cbs.dk
+45 3815 3130

ABSTRACT[1]
We present a new algorithm suitable for social recommender systems connecting web-clients based on their personal characteristics and interests. The algorithm has been tested in a web-based application with a conversational agent serving as a sophisticated switchboard operator. Based on a short client-agent conversation, the profile of the client is determined and compared to the profiles stored in a database. The current system is focused on the overall topic "ethical choices in life", demonstrating simple formal principles and computational tractability even within a notoriously difficult semantic domain. The algorithm is radically data-driven, informed by a background corpus of automatically collected text documents. The system is robust towards irregular input (misspellings, broken syntax), and this fact in combination with fine computational scaling properties could make it attractive for implementation in mass-services like Facebook and the blog-world. Freddy will be available for test runs at SRS-2010.

Author Keywords
Social Recommender Systems; Client matching; Man-machine interface; Ethics; Conversational agent

ACM Classification Keywords
Intelligent agents; Query processing; Relevance feedback; Information Search and Retrieval; Selection process

INTRODUCTION

Joseph Weizenbaum's classical program ELIZA [1] left the impression on many people that a machine could actually understand the psychological reality of a human.

Eliza:- Please state your problem.
Client:- I'm sad
Eliza:- I am sorry to hear that you are sad.
Client:- my wife just left me
Eliza:- Tell me more about your family.

Inspired by ELIZA, we developed a web-based conversational agent with a flair for ethics. The agent, Freddy, was intended as a sophisticated switchboard operator connecting web-clients of similar ethical inclination. The discourse planning system uses a simple psychological fact: you can learn a lot about your conversational partner by just exchanging a few remarks.

Freddy:- What do you do to improve the world?
Client:- I'm always riding my bike to work
Freddy:- Instead of taking your car?
Client:- yeah - better for the environment
Freddy:- What else do you do for the environment?
Client:- I use green energy, like wind generated electricity; I also use energy saving light bulbs
Freddy:- Do you wish to chat with another person who is also conscious of her carbon emission footprint?

The coherence in this conversation is based on a common ground concerning the ethical relevance and impact of the actions mentioned. In this case, Freddy discovered that "wind generated electricity" is associated with "carbon emission footprint" within the realm of *sustainability*. This kind of association and reasoning is not difficult for humans. Consider an example.

(1) I always use a push lawn mower
(2) Riding my bike to work
(3) Avoid dairy as much as possible
(4) Animal flesh is not for eating
(5) Only use green energy sources
(6) Do not kill and consume living creatures

People asked to label statements 1-6 with either **environmentalism** or **vegetarianism** according to their ethical motivation, uniformly pick (1), (2), and (5) as **e**, (3), (4), (6) as **v**. In spite of the huge linguistic and practical diversity of (1)-(6), people find the classification task easy.

Ethical terms that are easy to *apply*, may however be hard to *define*. Lexical definitions of ethical concepts tend to be expressed in other terms that are just as vague and abstract. By way of example, "solidarity" is explained by Wikipedia as "a unity of purpose or togetherness", and "social solidarity" as "the integration, and degree and type of integration, shown by a society or group with people and their neighbors". Such definitions seem next to impossible to fit in a rule-based reasoning system. Given our focus on ethically based match-making, we therefore opted for a

Workshop SRS'10, February 7, 2010 Hong Kong, China

data-driven (as opposed to rule-driven) approach. In our system, decision-making is thus based solely on statistical analyses of human actions and discourse as represented in a large background corpus.

In this paper we present those components of the system that are contributing to Freddy's ethical flair. A demo will available for test-runs at SRS-2010.

FREDDY'S VITAL PARTS

For readability, we use *ethical quality terms*, *EQTs* and *subsections of the Corpus* interchangeably. By a *type*, we refer to a lexical word form while a *token* is a particular occurrence of a type. We speak of *large* (*small*) *tokens*, meaning tokens with relatively many (few) occurrences.

The Corpus

The Corpus represents the system's world knowledge. It is subdivided in sections each representing a central theme. The global set of themes must (i) contain only very general terms, (ii) define the overall topics of the conversational realm. The determination of the theme set is the only non-automatic part of the preparation and maintenance of the client matching system. Currently, Freddy's conversational realm is defined by 23 terms (EQTs).

EQTs = {'solidarity' 'benevolence', 'sustainability', ...}

The current system uses a Corpus harvested with Google. Each EQT was installed as a search keyword. Visited html pages were stripped for everything but plain text. Long documents were pruned (global max = 5k tokens/doc), and all text lines preceding (following) the first (last) occurrence of the key word were skipped. Remaining tokens were normalized (non-alphabetic characters deleted, etc). Textlines with 2+ occurrences in a single EQT were only represented once in the Corpus (removing multiple copies of the same texts and paragraphs). Otherwise no filtering or token analysis was used.

Corpus samples

(from EQT "Solidarity"):

> "(..) The power structure of course is still controlled by white men. But the rise of a middle class of all races is real (..)"
>
> "(..) Leave the factory to go to meetings and demonstrations against the war strike against the state sponsored violence. Encourage your enlisted men in the armed forces to do the same (..)"

(from EQT "Veganism"):

> "(..) As far as the hot-dog and fanta comment Kelly made that is pointless unless she feels that all meat-eaters approach food this way (..)"
>
> "(..) What about school systems that get the green light for defining corn-syrup filled ketchup as a vegetable? (..)"

Most EQTs in the current Corpus consist of 100k to 300k tokens. The actual sizes and their relative differences have minimal impact on the ethical scorings, as long as each EQT is large enough to avoid sparse-data problems (more on this shortly). A practical lower limit is, say, 50k tokens per EQT.

The Actions database

The Corpus is supplemented by a database of actions, submitted by human clients on various occations. While the initial Actions Database consisted of 100 actions manually inserted by the author, since then Freddy has been accumulating the actions presented to him by his conversation partners. Currently the Actions Database contains more than 800 actions submitted by approximately 150 agents. Each action is stored with a reference to the originating agent and a canned ethical profile (see next section) for computational efficiency. The Actions Database is used by Freddy for match-making among his human conversational partners based on their ethical profiles.

The ethical profile

Freddy's reasoning is based on a central data-structure called the ethical profile (EthPro), a function mapping a list of input tokens *I* (typically a description of an action) onto a set of pairs <*Q*,*V*>, where *Q* is an EQT and *V* the related score value (a positive real number).

(A) EthPro("eat less meat") =

Vegetarianism	1.243
Veganism	0.954
Animal rights	0.863
Kindness to animals	0.734
Environmentalism	0.583
Spirituality	0.454
(...)	

(B) EthPro('disobey your rulers") =

Obedience	0.560
Anarchism	0.373
Strength of character	0.286
Privacy	0.253
Humbleness	0.201
Responsibility	0.199
(...)	

As shown in (A) and (B), EthPros are typically sorted by *V*.

Observe in (B) that "disobey your rulers" picks Obedience as its most salient EQT. This proposition of course does not exemplify obedience, but the opposite. This illustrates a feature of the Ethical Scoring regime, the *topic* of Obedience being highly relevant in discourse on attitudes towards rulers. Similarly, "meat" and "vegetables" show an affinity to Vegetarianism while food items like "eggs" and "fish" score relatively higher for Veganism (reflecting the distinguishing food items in the particular diets).

Ethical score values are computed as the product of two factors, Global Heterogeneity (GloHet) and Local Homogeneity (LocHom). GloHet has a high value for types unevenly distributed over the Corpus as a whole, for instance when most occurrences are concentrated within a few subsections. GloHet thus measures the capacity of a type as an ethical discriminator.

LocHom measures the distribution of a type *within* a EQT. A high LocHom value for token T in section Q indicates that T is equally frequent in many or all documents in Q, in other words that T belongs to the common vocabulary of this particular EQT.[2]

$$\mathrm{GloHet}_C(T)=\frac{SDev(T,C)}{Mean(T,C)}$$

$$\mathrm{LocHom}_C(T,Q)=1/\frac{SDev(T,Q)}{Mean(T,Q)}$$

$$\mathrm{EthSco}_C(I,Q)=\frac{\sum_{i=1}^{|I|}\mathrm{GloHet}_C(I_i)*\mathrm{LocHom}_C(I_i,Q)}{|I|}$$

T is a token; I is a string of tokens I_1, I_2, .., $I_{|I|}$ (e.g. a proposition); C is a corpus with sections c_1, c_2, .., c_N representing the EQTs; each c_x consists of a set of documents $d_{x,1}$, $d_{x,2}$, .., $d_{x,M}$; $SDev(T,c_x)$ is the standard deviation of frequency values $Freq(T,d_{x,1})$, $Freq(T,d_{x,2})$, .., $Freq(T,d_{x,M})$ while $Mean$ is the average of the same values. The $Mean$ part of the formulae serve to relativize the variance values making small and large tokens numerically comparable.

The relative difference between two ethical profiles is computed by simply summing up the differences between the individual EthSco values.

$$\mathrm{Diff}(EP,EP')=\frac{\sum_{i=1}^{|Q|}\mathrm{ABS}(V_i-V_i')}{|Q|}$$

EP and EP' are ethical profiles both defined on a domain Q of EQTs: $\{<Q_1,V_1>, <Q_2,V_2>, .. ,<Q_{|Q|},V_{|Q|}>\}$ and $\{<Q_1,V_1'>,<Q_2,V_2'>, .. ,<Q_{|Q|},V_{|Q|}'>\}$, respectively.[3] The Diff formula plays a central role in the action matching and agent matching procedures (see fig. 1). Profile matching is computationally cheap, and agent profiles can be computed off-line, ensuring scalability and short response times.

Other EthSco and Diff formulae can be conceived, and many have been considered (see note 2). The present version is the answer to a number of desiderata. EthSco calculation is invariant (everything else being equal) to:

- *input length* (example: "save on oil" has the same ethical profile as "save on oil, save on oil, save on oil")
- *token count* (example: types "social", "fair", and "opposition" have approximately same GloHet value while they are very different in size: 841ppm, 116ppm, and 47ppm)
- *absolute size of EQTs* (example: doubling each EQT in C does not change the EthPro for a fixed I)
- *relative size of EQTs* (ex: doubling each document in a EQT does not change the LocHom value for a fixed I)
- *corpus composition* (ex: profiles with distinct EQT composition are directly numerically comparable, given a non-empty intersection),

These invariances are desirable, making corpus maintenance much more flexible. One can add texts to the Corpus continuously, even to selected EQTs only, and still maintain backwards compatibility. Also, the ethical profiles stored in the Actions Database do not have to be refreshed each time a new document - or even a new EQT - is introduced in the Corpus.

Discourse planning

When invoked by a client, Freddy presents himself and then continue to produce a prompting question such as "What actions do you take to improve the world?".

The client's answer is scored using EthSco, and the derived ethical profile is used for action matching in A (see fig. 1). This process returns two lists, (i) all actions in A and, (ii) all clients in A, sorted by EthPro proximity[4]. These lists are passed on to the Discourse planning module (essentially an Eliza-style text generator, presented here by an example only). As the generated reply depends directly on the set of actions in A, the system responses change over time as more and more Client submissions are added.

Client:- I'm always riding my bike to work
Freddy:- Instead of taking the car?

In this case, Freddy found a near-match in A:

"I take my bike to work instead of taking the car"

The cut'n'paste procedure relies on fairly superficial word string comparisons, but with ethical profile matching as a relevance control regime.

I take [my bike to work] instead of taking the car
I'm always riding [my bike to work]

"I take" and "I'm always riding" only contain types with low GloHet (interpreted as a license to ignore) while the final phrase "instead .. car" has an EthPro closely matching that of "my bike to work" (interpreted as a license to quote).

Text bits from the Corpus may be used as fillers in the same manner when no adequate match can be found in A.

When Freddy has gathered a few client actions, the exit point is reached where Freddy offers a link to a discussion partner (again based on ethical profile comparisons).

Freddy:- Do you wish to chat with another person who is also conscious of her carbon emission footprint?

The user may answer 'yes' or choose to continue the conversation with Freddy.

[2]Our early EthSco formulae did not incorporate the EQT internal distribution (LocHom), leading to an unwanted bias e.g. in cases where a single document had a large number of occurrences of a particular type absent from the other documents in the EQT and the Corpus in general.

[3]This Diff formula takes EP and EP' to be defined for the same Qs; this requirement is easily relaxed so that even profiles defined for distinct EQT sets are comparable as long as there is a (preferably substantial) overlap.

[4]Ethical profiles for agents (represented as sets of actions) can be computed in two ways: either by pooling all tokens produced by an agent calculating a single profile, or by scoring each action separately and then calculate the mean of all profiles. We are currently testing both methods.

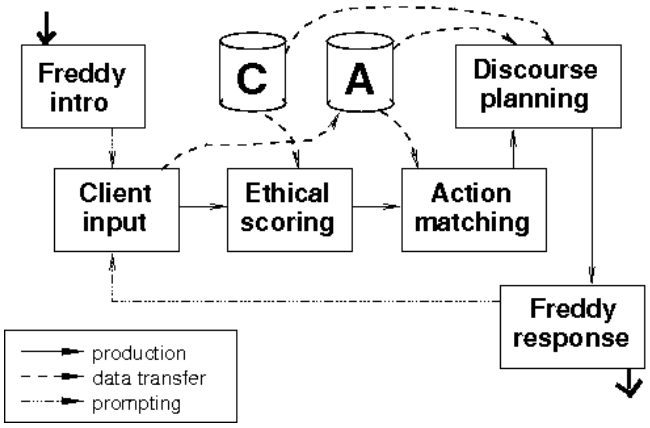

Figure 1. Functional modules
(C = Corpus database, A = Actions database)

EVALUATION

In a preliminary evaluation session, Freddy was tested by 8 adult users all fluent in English. Each user was given three "ethical personalities" picked from the global EQT domain, in turn answering Freddy's questions from each EQT perspective. Freddy's final decisions ("Freddy response" in fig.1) were recorded and later evaluated in a blind review. The procedure was: User U_1 receives EQTs *E1*, *E2*, and *E3*. Freddy selects partners *P1*, *P2*, *P3*, respectively. Users U_2 through U_8 are presented with *E1*, *E2*, *E3* and *P1, P2, P3* in random order and asked to pair them (they are allowed to consult all actions of *P1, P2, P3* in the Actions Database). Same procedure is repeated for all users, resulting in 21 *E-P* pairs for each user, or 168 pairs in total.

A random pairing gets 33.3% correct pairs on average. In contrast, Freddy got 81% correct pairs (136 of 168). In spite of his formal simplicity, Freddy seems to make reasonable decisions within a very sophisticated knowledge domain.

More elaborate test sessions are in preparation, including user ratings of the interaction with the system.

DISCUSSION

In the proposed client classification system, property definitions are replaced by structured data repositories. Whereas lexicographers recommend homogeneous denotation principles, our document collection benefits from including as many independent, heterogeneous text sources as possible, blogs, Twitter, Google search, interest groups - anything *but* traditional formal explanations. This way the plurality of term uses will be maximally covered.

Morphological and syntactic analysis (lemmatizing, stemming, PoS tagging, parsing) is avoided in the current system. Types like "obey", "obeys", "obeyed", "obedience", "obedient" are treated as independent atoms, as are "I", "me", "my", "we", etc. Most Q/A systems employ at least some linguistic parsing (e.g. [1], [5], [6]) in order to support compositional-semantic decoding in structured knowledge fields such as appointment management ([2]), information retrieval from web pages ([3]), and tutorial services ([4]).

From a developer's point of view, data-driven methods however have several advantages.

- language independent input analysis procedure
- no need for labour intensive rule writing
- robust to unknown words, misspellings, broken syntax
- the system performance is improving over time, new input being added to the existing database

In many information retrieval systems, semantically empty words like pronouns, copula verbs, and prepositions are ignored. However, we did some experiments showing that even oppositions like "we" and "I", expected to be ethically neutral, actually do contribute to the distinctive powers of Freddy. By way of an example, the *I*:*we* ratio is more than four times higher in Veganism (a personal goal) than in Environmentalism (a social goal). Even if vegetarian and environmentalist goals are often overlapping - such as reducing the consumption of animal protein - Freddy is able to discover the subtle ethical difference between "I should eat less meat" and "we should eat less meat". We therefore do not edit the input string before computing its EthPro.

A similar discussion concerns the syntactic structure of the user input. The current Freddy treats the user input as a bag of words with no intrinsic ordering. Most designers of dialogue systems would probably argue that such an architecture loses useful information. This is undoubtedly true for highly controlled input, but Freddy is intended as an informal chat partner and must be prepared for input with erroneous spelling, broken syntax, and even deliberate chat-style reductions making rule-based methods vulnerable. Still some shallow syntax parsing may be rewarding; we are currently looking into this possibility.

The presented algorithm is still in the making, but we do have a first conclusion: People like to chat with Freddy. We suggest that *conversation-driven client matching* should play an active role in next-generation social recommender systems offering the client an experience of personal service.

REFERENCES

1. Weizenbaum, J. Eliza - a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communic. of the ACM 9*, 1 (1966), 36-45.
2. Myers, K. et al. An Intelligent Personal Assistant for Task and Time Management. *AI Magazine 28*, 2 (2007)
3. Kyoung-Min, K. An Intelligent Conversational Agent as the Web Virtual Representative Using Semantic Bayesian Networks. *PRICAI* (2006)
4. Robinson, S. et al. What would you ask a Conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting. *LREC* (2008)
5. DeVault, D. et al. Making Grammar-Based Generation Easier to Deploy in Dialogue Systems. *9th SIGdial Workshop on Discourse and Dialogue* (2008)
6. Gandhe, S. et al. Improving Question-Answering with Linking Dialogues. *IUI'06* (2006)