ACT-R Model for Credibility Judgments of Micro-blogging Web Pages

Q. Vera Liao (liao28@Illinois.edu)
Department of Computer Science, University of Illinois, 201 N. Goodwin Ave
Urbana, IL 61801 USA

Peter Pirolli (pirolli@parc.com)
Palo Alto Research Center, 3333 Coyote Hill Rd.
Palo Alto, CA 94304

Wai-Tat Fu (wfu@illinois.edu)
Department of Computer Science, University of Illinois, 201 N. Goodwin Ave
Urbana, IL 61801 USA

Abstract

In this paper, we propose an ACT-R cognitive model for making credibility judgments about the credibility of Twitter authors. We abstracted the cognitive processes involved in three levels: attending to information on Web page, comprehending information to identify credibility cues, and integrating credibility cues to make a judgment. We represent basic knowledge required for credibility judgment using declarative memory in ACT-R that is seeded with experiences of Twitter messages that have been passed through a Latent Dirichlet Allocation topic modeling process. Comparisons of model credibility judgments to human credibility judgments from controlled experiments show weak to strong correlations that range from $r = 0.32$ to $r = 0.75$ depending on the specific task.

Keywords: Web credibility judgment, ACT-R

When people make credibility judgments about Web-based content its sources, people must perceive, comprehend and deliberate on the merits and flaws of available cues to make the judgment. Complexity arises from the fact that the judgment is rarely based on a single cue, but requires the integration of multiple cues. These cues may interact with or contradict each other, and accumulate over the course of interaction with the Web content. We present a cognitive modeling approach to investigate multi-cue Web credibility judgment.

Cognitive models have been applied to explain and predict human interaction with Web-based content, primarily focusing on relevance-based browsing or search. For example, MESA (Miller & Remington, 2004) and SNIF-ACT (Fu & Pirolli, 2007) are models that simulate how users navigate through websites to search for information relevant to a given task. Web credibility judgment is a complex high-level cognitive process that may be highly dependent on the goal of the user. Therefore, instead of building a universal model, our goal is to propose a framework, or a methodology that can be easily modified for different contexts, and demonstrate it with a specific task. In this study, we attempt to build an ACT-R model of credibility judgment when processing Twitter micro-blogging content,.

Website credibility models are often conceptualized along two dimensions. One dimension, represented by stage models (Wathen & Burkell, 2002), focuses on the iterative process of credibility evaluation, i.e., how the assessment takes place when users open a page, read the contents, and are further involved with the site. The other dimension, following a bottom-up approach, seeks to examine what elements on a Web page, and to what extent, impact users' credibility judgments. Detailed cognitive models have the potential to model the iterative processes of stage models and the impact of specific Web cues in different task and content contexts.

We chose to analyze a task with simplified Twitter page, which allows us to ignore the complex interactions between multiple types of information cues but focus on the iterative process of attending to, processing and evaluating information on a Web page. This study was also motivated by the potential value of building predictive models for evaluating information credibility of micro-blogging, and more broadly, user generated contents on Internet.

In the following section, we will first introduce the modeling task and a preliminary study conducted with the task. Conclusions drawn from the preliminary study are incorporated into the ACT-R model. In the second part we will describe the ACT-R model. Lastly, we will present a model validated it by a human data by a second experiment with the same credibility judgment tasks.

Modeling Task and Preliminary Study

The modeling task was based on a Twitter study conducted by Canini et al.(2011). Twitter is the popular micro-blogging service that enables its users to add text-based posts of up to 140 characters, known as "tweets", on their own page. The goal of the study was to explore what factors on a Twitter page may impact users' credibility judgment about the Twitter author. Understanding this process is important because it may help improve the design of micro-blogger recommendation systems and user interfaces to help users to discover credible sources and content.

In the Canini et al. (2011) experiment, participants were presented with a page generated to represent individual

Twitter users. Each of these generated pages, included a user name and icon, a set of social status statistics (number of following, followers and tweets), 40 tweets by the user, and a word cloud summarizing the Twitter user's generated content (Figure 1). Among other things, each participant was asked to rate presented Twitter users' credibility in making judgments in the specific domain of car purchases. Three variables were manipulated in Canini et al (2011) in constructing the Twitter pages representing users:

(1) *Content domain*. The top 10 experts suggested in the WeFollow directories of *car*, *investing*, *wine*, *fantasy football*, *dating* plus 10 random accounts were selected. WeFollow is a popular Twitter user recommendation system. It has topic directories such as car, football, etc, where users can sign up if they are experts or interested in the topic. Wefollow ranks all users based how many users in the same directory are following him/her.. Experts from the car domain were considered *on-topic* with respect to the target task of judging recommendations for car purchases, the other domains were *cross-topic*.

(2) *Social status*. For each page, the social status was randomly set to be high or low. For a high social status, the presented user had a large number of following/followers (more than 1000) and a large number of tweets (more than 100).

(3) *Visualization*. The page was randomly set to be tweets only, word cloud+tweets, and word cloud only.

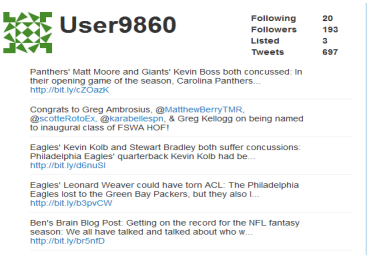

Figure 1. Modeling Task Interface

The Canini et al (2011) results showed that the directory from which the Twitter author was selected had strong influence on perceived credibility. Not surprisingly, those selected from the car directory (on-topic) led to significantly higher credibility ratings than those from other directories (cross-topic). It was also found that users considered someone who talked a lot about dating were the least credible in giving car price suggestion, while experts in investing had a credibility rating in between the dating and car directories, possibly because the task of suggesting car price is related to financial decisions. It was also found that social status and visualization factors had smaller but statistically significant influences on credibility judgment.

We built an ACT-R model for this credibility judgment task. The credibility ratings given by the model are positively influenced by on-topic contents and negatively influenced by certain cross-topic contents. The model also has the capacity to process other contextual features on the Web page, such as social status.

Model Framework

We now present the general framework of the cognitive model for Web credibility judgment, and how this is implemented in ACT-R. Representations of knowledge are stored in declarative and procedural memory modules in ACT-R. Declarative memory, consisting of facts such as "BMW is a car brand", is represented by memory chunks built into the model. Procedural memory, representing knowledge about how we do things, e.g., how to judge if an information source is credible, is represented as productions.

As shown in Figure 2, the model framework assumes a process consisting of three phases. First, the model attends to information on the page. In the first phase includes processes that mostly involve perception and attention, such as fixing attention on Tweets and initiating reading. For the ACT-R model, by attending to a Tweet, e.g., "happy driving and car shopping", the model will recognize the word "happy", "driving", "car" and "shopping" by making use of its vocabulary knowledge in declarative memory.

In the second phase,, the model comprehends information it has attended to, which leads to the identification of information cues that may potentially impact the credibility judgment. We use the spreading activation mechanism of ACT-R to implement this process. Retrieval of each chunk in declarative memory in ACT-R is determined by a chunk's activation. Activation reflects the degree to which a chunk is likely to be needed or relevant in the current context based on prior experience. The chunk with highest activation and above a set threshold is most likely to be retrieved. In addition to the base level activation which reflects the prior use of the chunk itself, the chunk will also receive activation spread from related chunks currently attended by the model.. For example, when the model reads the Tweet "happy driving and car shopping", each of the word spreads activation to potentially related topics. Both the word "car" and "driving" spread activation to the "car" topic, making its activation higher than other topics, e.g., "shop", which only receives activation from the word "shopping". Then the topic "car" will be retrieved, as being identified to be the topic of this particular Tweet. Optionally, this phase may also involve inferences made based on the perception of other features on the Website. For example, if the model reads a large number of followers, it may identify it as a cue of high social status.

In the third phase, the model will deliberate on the information cues it identified and integrated them to make a credibility judgment. In the ACT-R model, we use the *blending mechanism* (***ref?***) to implement this phase. When using blending, if there are multiple candidate chunks satisfying the retrieval request but with different values in certain slots of the different retrieved chunks, the model will construct a chunk that contains slot values that "blend" over those multiple values. More specifically, ACT-R will

retrieve a chunk that contains a compromise value, V, in the target slot that is determined by:

$$V = Min\sum_i P(1 - Sim(V, V_i))^2$$

where Vi is the value held in the target slot of the existing chunks i. Pi is the probability of retrieving existing chunk i, which is determined by the activation of chunk i. When making a credibility judgment, we assume that the model utilizes knowledge of previously stored instances of credibility judgments, i.e., it has prior knowledge that a certain cue is an indication of being credible or non-credible for, and strength of that indication varies. The model blends all the instances it retrieves based on cues from perceived Web content to make the judgment. For example, the model will identify that topics "car", "gas" and "dating" are discussed in the Tweets. It will then decide that "car" is a strong indicator of credibility for giving car price suggestion, which is represented by a strong activation spread from chunk "car" to chunk "credible". Similarly, it may decide "gas" is a less strong indicator of credibility. However, "dating" may be an indicator of non-credibility and thus spread activation to the chunk "non-credible". Then the model will integrate the credibility indications of all cues according to the total activation received by the credible chunk and non-credible chunk to make the credibility judgment.

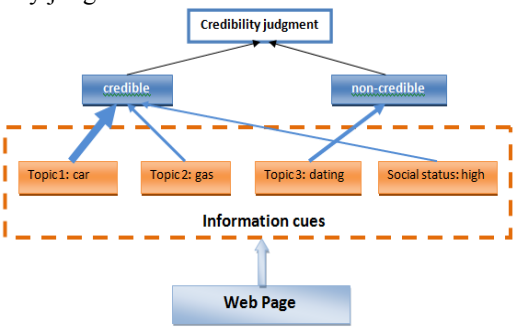

Figure 2. Model Framework

ACT-R Model for Twitter Author Credibility Judgment

The ACT-R model for Twitter page credibility judgment uses two buffers in addition to the basic ACT-R buffers: a word buffer and a credibility cue buffer. The content of the word buffer reflects the text that the model attends to and holds in a short-term memory. The credibility cue buffer contains cues identified by the model which may potentially have impact on credibility judgment. In the following section I will describe how we construct the declarative and procedural memory to work with the two buffers.

Declarative Memory

The declarative memory of this ACT-R includes word chunks, topic chunks and credibility chunks, and optionally, contextual cue chunks. Because the Web credibility judgment process may involve frequent use of declarative knowledge, it is important to build declarative memory that allows adequate knowledge for such process. Therefore, to enable the model to process Twitter pages, we built a corpus by collecting all tweets from 1800 twitter accounts randomly chosen from different WeFollow directories, and constructed the declarative memory from this large dataset.

Word Chunk

We identified the 3000 stemmed words (which are not stop words such as a, the, of, etc) with the highest frequency from the Tweets corpus. Word chunks to represent each of the 3000 words were added into the declarative memory. These represent the vocabulary knowledge the model has to process the contents.

Topic Chunk

We used Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) to identify topics that can be used to comprehend Twitter message content. LDA is a generative model which posits that a document, i.e., the collection of observed words, is a mixture of unobserved topics and that each word's creation is attributed to one or several of the document's topics. We exploited an LDA topic model produced in Canini et al. (2011) that used documents that were constructed by aggregating all the retrievable tweets produced by individual WeFollow users (maximum = 3000 tweets). Following Canini et al. (2011), we selected 500 topics with the highest frequency to be the topic chunks in declarative memory. They represent the knowledge for processing and comprehending Tweets. Each word chunk is associated with one or multiple topics.

Contextual Information Cue Chunk

All the contextual information cues, if any, could be added into declarative memory as cue chunks. For example, to process social status in the task, we could add "high social status chunk" and "low social status chunk" into the declarative memory.

Credibility Chunk

We built two credibility chunks, a "credible" chunk and a "non-credible" chunk which have a value slot to represent the two extreme values (rating 1 and rating 7) of credibility judgment ratings. Each credibility cue chunk (including topic chunk and contextual information cue chunk) is associated with either a credible chunk or non-credible chunk.

Procedural Memory

The procedural memory was built to execute the credibility judgment process as shown in Table 1. The model will start by reading the textual content in sequence (i.e., from left to right, top to bottom). When the model attends to a word, and it has a corresponding word chunk in the declarative memory, the chunk will be retrieved and placed in the word buffer. With the limitation of short term memory, only a limited number of words will be stored in the buffer. When the word buffer reaches its capacity, if a new word chunk is retrieved, the earliest word attended will be removed, and each existing cue in the buffer will be

moved to the earlier slot. Hence the model will iteratively hold the latest words it attends to in the word buffer.

When processing the contents, the model attempts to identify topics based on what it has just read. At any moment, the word buffer contains a list of words. Each of the word chunks is associated with one or multiple topic chunks in the declarative memory. All these words will collectively decide the strength of association spreading to the topic chunks. The topic that is above retrieval threshold and receives highest activation will be placed into the credibility cue buffer. Since the list of words in the word buffer will continuously change, the model may identify multiple topics as the model reads through the page. For the current model, we only allow topics that are not currently in the credibility cue buffer to be retrieved. Optionally, the credibility cue buffer has slots to hold contextual credibility cues. The credibility cue buffer also has limited number of slots, and will only keep the latest credibility cues.

Resembling human behavior, the model may stop before it finishes processing all infornation. Anytime the model identifies a new credibility cue, it chooses between the production that halts further reading and a production to continuing processing. In ACT-R, when there are multiple productions wating to be fired, the chances that production *i* will be fired is decided by:

$$P(i) = \frac{e^{U_i/\sqrt{2}s}}{\sum_j e^{U_j/\sqrt{2}s}}$$

where U_i represents the utility value set for production *i* and *s* is a utility noise parameter,. We set the utility of the production for continuing processing to be higher than the production to halt reading. Therefore at different points of processing the Web content, the model has chance to stop, but the chance is still lower than that of continuing reading..

When either the model chooses to stop or it reaches the end of the page, the production for making the credibility judgment will be fired. As discussed in the previous section, there is a credible chunk with a rating slot of value 7, and a non-credible chunk with a rating slot of value 1. They receive activation spread from the credibility cue buffer, as positive credibility cues are associated with the credible chunk, and negative credibility cues are associated with the non-credible chunk. The model uses the blending mechanism to blend the rating values of credibility chunk and non-credibility chunk based on the activation of the two chunks.

Table 1. Model Procedural

Attend to word IF there is corresponding chunk in declarative memory THEN push the chunk into word buffer	IF NOT THEN attend to next word
Hold word in word buffer IF there is open slot in word buffer THEN hold the word chunk in the latest open slot	IF NOT THEN remove the earliest word and move each word chunk to an earlier slot to open the latest slot
Understand topic IF there is topic(s) above retrieve threshold & the topic(s) is not held in the credibility cue buffer THEN retrieve a topic	IF NOT THEN attend to next word
Hold topic in credibility cue buffer IF there is open slot in credibility cue buffer THEN hold the topic in credibility cue buffer	IF NOT THEN remove the earliest cue and move each cue to an earlier slot to open up the latest slot
Decide to stop of continue IF stop production is fired THEN start to make credibility judgment	IF NOT THEN attend to next word
Make credibility judgment IF model stops reading or no more content left for processing THEN make credibility judgment blending credibility chunks	

Strength of Association

ACT-R calculates the activation of each chunk by:

$$A_i = B_i + \sum_k \sum_j W_{kj} S_{ji} + \varepsilon$$

where B_i is the base-level activation, which reflects the recency and frequency of practice of chunk *i*. The component $W_{kj}S_{ji}$ reflects spreading of activation from retrieved chunks to related chunks in the declarative memory. *S* represents the strength of association. *W* can be set to decide the weighting of different slots in a buffer to spread activation to the declarative memory. ε is the system noise value.

There are two phases in the model where the activation spreading plays a role: 1) the emergence of topic is determined by the collective activation spread from the words held in word buffer, and 2) the activation of credibility chunks is determined by the collective activation spread from the credibility cues held in the credibility cue buffer. We will describe the rules we used to set the strength of spreading activation below.

Strength of association from word to topic

By using the LDA topic model for the Twitter corpus described above, we calculate the strength of association from word to topic by:

$$S_{wt} = \log(P(w \mid t) / P(w))$$

where $P(w \mid t)$ is the LDA-estimated probability of word ω given the occurrence of topic *t* and $P(w)$ is an estimated of the probability of word occurrence.

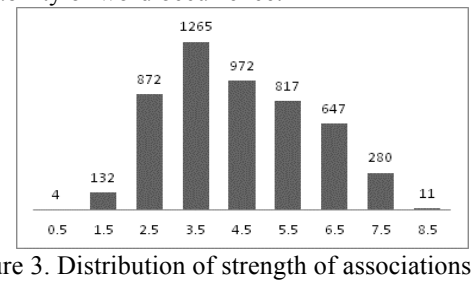

Figure 3. Distribution of strength of associations from word to topic

For the model, we set the limit of number of word slots for each topic chunk to be 10. It means we only identify the strength of association of the top 10 words for each topic, and overall we identified 5000 strength of associations (10

for each of the 500 topics). The distribution of the strength of association (number of associations falling in each range of strength) is shown in Figure 3. This approach enables the model to have the knowledge to infer the potential explanations (i.e., topics) of each word that it attends to.

Strength of association from credibility cue to credibility

Strength of association from topic chunks to credibility chunks indicates the extent to which the particular topic is regarded as an indicator of credibility or non-credibility by the model. The model reads the task description and attends to key words of the task (e.g., for the car price suggestion task, the key words are "car" and "price"). For each of the key words, the model attempts to identify topics that are highly related to the key words. We set the current model to select the top 30 topics, with which each key word chunk has the highest strengths of association. Then the model increases the strength of association from the topic to credibility chunk by the same amount of strength. It allows the model to use a bottom up approach to identify topics that are associated with the task goal and that may have positive impact on credibility judgment.

According to the results of our preliminary study, there seemed to be topics with negative effects on the credibility rating (e.g., dating related topics). While it is difficult to exhaustively identify all the negatively associated topics, since we only intend to test the model with directories of car, dating and investing at the current stage, we manually selected a few topics that are strongly associated with words frequently used by authors in dating directory (e.g., dating, sex, etc), and set strength of associations from these negative topics to the non-credible chunk.

Similarly, contextual cue chunks in the credibility cue buffer, if any, will spread activation to either of the two credibility chunks. For example, the high social status chunk, if held in credibility cue buffer, will spread activation to credible chunk. The weighting of slots for different types of credibility cue can be set according to the task context.

Pilot Validation

We used the same setup and procedure as in the Canini et al. (2011) experiment, which asks participants to rate a Twitter author's credibility for giving car price suggestions. However, instead of manipulating multiple features on the page, we focused on only users' tweet contents. We selected the latest 40 tweets from the top 10 users recommended in the WeFollow directories for cars, investing and dating. We recruited N = 7 participants to complete the credibility rating task. Each participant judged all the 30 pages in random order.

We first performed a repeated measure ANOVA on participants' credibility ratings, with author domain (car, dating, investing) as the independent variable. The result showed that the main effects of directory is significant ($F(2,12)=4.82$, $p=0.03$), meaning credibility ratings given to the authors from the three directories are different. Then we compared each pair of author directory. It showed that the ratings given to authors from car directory are significantly higher than those from dating directory ($F(1,6)=12.05$, $p=0.01$), and marginally significantly higher than those from investigating directory ($F(1,6)=3.98$, $p=0.09$).The model results showed the same pattern. As the model results may vary if it stops reading at different parts of the page, we ran the model for 10 times and calculated the mean ratings for each page. We performed t-test between each pair of author directories for model results. It shows the ratings given to Twitter author selected from car directory are significantly higher than those from dating directory ($p<0.01$), and those from investing directory ($p<0.01$). The results suggest that, the model, like human participants, is able to infer the source credibility for the task goal (i.e., car price suggestion) based on the micro-blogging content created by the person.

We are aware that the perceived credibility varies even for Twitter authors selected from the same directory. For example, some car experts may not necessarily talk about cars in their Tweets, while others may tweet about it frequently. Potentially, one practical use of a cognitive model for Web credibility judgment is the capability of predicting perceived credibility for individual pages. We therefore looked into the correlations between human judgment and model judgment for individual pages. Specifically, we expect the model to be able to differentiate higher credibility from lower credibility Twitter sources as judged by humans.

Figure 4 shows the human results and model results for credulity ratings about 10 users chosen from the WeFollow directories of cars, investing and dating. The fit between human and model results for car directory is R^2=0.56, correlation for investing directory is R^2=0.30, correlation for dating directory is R^2=0.10. Although the results do not show a good fit for investing and dating directory, we are aware that the current model may not be able to exhaustively identify information cues that negatively affect

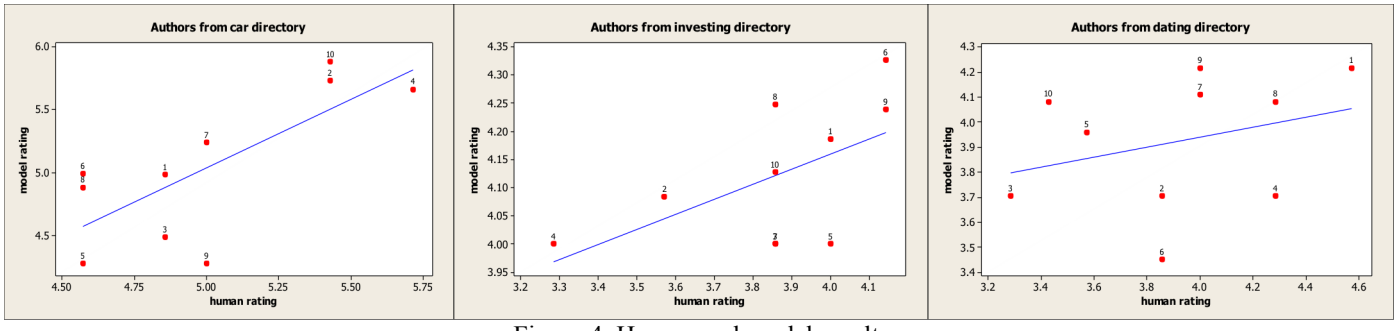

Figure 4. Human and model results

credibility judgments.

At a broader level of analysis we tested to what extent the model could predict at the valence (i.e., low vs high) of the credibility judgment. To this end, for the 10 pages with authors from the car directory, we performed a median split analysis of Twitter user credibility rating. Each Twitter user was coded as being (1) high-credibility or low-credibility baedon whether it was above median or below median in terms of avergage human rating and (2) high-credibility or low-credibility based on whether the Twitter user was above median or below median on model ratings. The results showed that, for 8 out of 10 pages, human results and model results fall into the same bucket (with the exception of 1 high and 1 low credibility page). To further verify these pages are perceived to have different valence of credibility, we performed repeated measure ANOVA with human ratings for the 8 pages, with the valence (high/.low) as independent variable. It shows the ratings are significantly different ($F(1,6)=10.52$, $p=0.02$). We performed the same analysis for authors selected from investing directory. We also found, for 8 out of 10 pages, human ratings and model ratings fall into the same high or low bucket (with the exception of 1 high and 1 low credibility page). The ANOVA verified the ratings given to the two groups of pages is marginally significant ($F(1,6)=4.52$, $p=0.07$). We did not look into the dating directory because of the lack of knowledge about negative cues as discussed earlier. These results proved that the model was able to predict the valence of credibility for individual pages.

Discussion

In this study, we proposed a framework for a cognitive model for making credibility judgments of Web content or its sources, and implemented it in ACT-R. We exploited Twitter content to induce an LDA topic model that was used to seed declarative memory and support an instance-based judgment process based on the ACT-R blending mechanism. In general, the model is able to infer the level of credibility of Twitter authors by differentiating authors with on-topic content for the task goal and those without. It is also able to predict the perceived credibility of individual users with on-topic contents.

The model performs three phases of cognitive processing to make a credibility judgment of Web content or sources: attending to information on the page, comprehending the information to infer credibility cues, and making credibility judgmenst by integrating these credibility cues. During the comprehending phase, the spreading activation mechanism of ACT-R is used to identify the most likely explanation when there are multiple pieces of observed information and each may have multiple explanations. The blending mechanism is used to generate a judgment by integrating credibility cues, each of which may indicate a different level of credibility. Although we built the model with a Twitter author judgment task in this paper, by changing the model knowledge for processing information on a Web page, and knowledge about credibility of different cues, the model could be modified to apply to different media, content, or sources.

The major limitation of current model is its lack of complete knowledge about the credibility indications of various information cues, especially those that may negatively impact credibility judgments. Future research is needed to explore this research question.

Reference

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22, 261-29

Anderson, J. R. & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum

Blei,D. M., Ng, A.., Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3 (4–5), 993–1022.

Canini, K. R., Suh, B., & Pirolli, P. L. (2011). Finding credible information sources in social networks based on content and social structure. *In Proceedings of the third IEEE International Conference on Social Computing (SocialCom).*

Fu, W.-T., Pirolli, P. (2007), SNIF-ACT: A Model of Information-Seeking Behavior in the World Wide Web. *Human-Computer Interaction*, 22, 355-412.

Lebiere, C. (1999). A blending process for aggregate retrievals. *In Proceedings of the 6th ACT-R Workshop.* George Mason University, Fairfax, Va.

Miller, C. S., & Remington, R. W. (2004). Modeling information navigation: Implications for information architecture. *Human Computer Interaction*, 19, 225–271

Wathen, C.N. and J. Burkell(2002), Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology* , 2 , 134-144.