18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Deriving Agent-Based Simulation Models from Textual Data

Dekker, A.H.

Defence Science and Technology Organisation, Australia
Email: dekker@acm.org

Abstract: This paper describes a project to investigate the feasibility of building agent-based social simulation models from a combination of textual data and limited observational notes.

As an example, we have developed a proof-of-concept Java-based simulation of one notional day at a scientific conference, based on the MODSIM 2005 International Congress on Modelling and Simulation.

Textual processing of the conference proceedings was used to build a common-interest network between authors, which in turn was used to drive agent processes for choosing paper presentations to attend and people with which to interact. Specifically, correlations between frequencies of word pairs was used to derive similarities between papers, and hence between authors.

Diary notes taken during the MODSIM 2005 International Congress were also used in developing the simulation, which was written in the Java programming language.

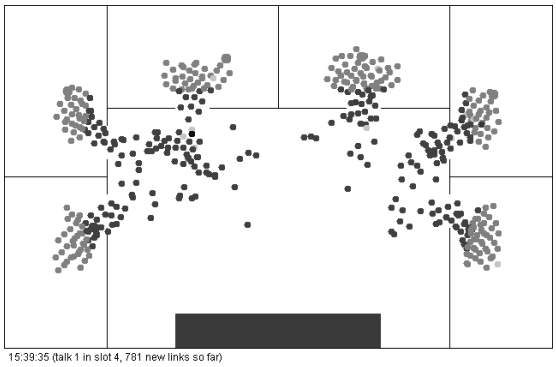

Figure (i). Conference simulation: participants leave the conference foyer area to attend paper presentations in seminar rooms.

The conference simulation shows realistic emergent phenomena, such as increasing clustering of conversing participants during the course of the day.

The use of textual data does not completely eliminate the need for observational data, and more extensive observational notes would have assisted in producing a more detailed simulation of MODSIM 2005. However, we have demonstrated that the analysis of textual data can be used to simplify model development.

In developing more general organisational simulations, time lags in the production of streams of text documents can be used to estimate the delays inherent in organisational processes. As a preliminary investigation of this approach, which we intend to pursue in further work, we show how a time delay can be extracted from the simple text snippets contained in two Internet news feeds.

Keywords: *social network, text processing, agent-based simulation, organisational simulation*

1. INTRODUCTION

The goal of the work described in this paper is to investigate the feasibility of building agent-based social simulation models from a combination of textual data and limited observational notes. This is intended to help produce predictive simulation models of organisations for which more detailed observational data is unavailable or difficult to obtain.

In the work presented here, we simulated one notional day of a scientific conference (MODSIM) attended by the present author in 2005. The textual data used was the electronic conference proceedings (Zerger and Argent, 2005), in which filenames are based on the name of the primary author. This electronic data was supplemented by diary notes made during the conference. Readers are invited to compare this work against their own conference experiences.

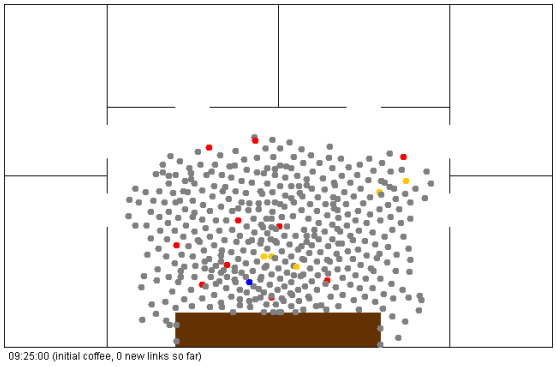

Figure 1. Conference simulation: initial coffee session. The blue agent represents the present author, red and yellow agents are participants with similar interests to the present author (yellow where an actual social link exists), and grey agents are other participants.

The textual data from the conference proceedings was used to build a network of common interest between authors. This represents a *potential* social network of researchers who would benefit from communicating with one another. Within the simulation model, this interest network determines which paper session attendees will go to, and whether they will ask questions of the presenter. As participants get to know each other through these interactions, potential social network links become *actual* social network links, which influence the subsequent behaviour of participants. Some of the potential social network links were deemed to be actual at the beginning of the simulated conference day.

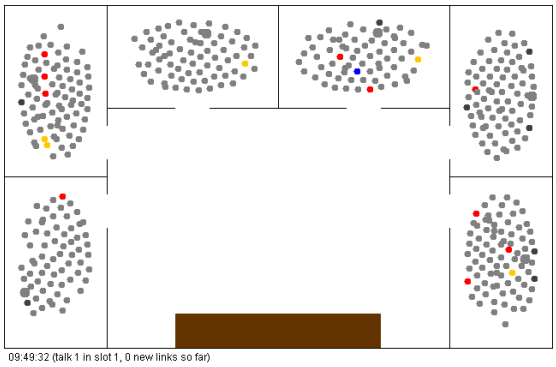

Figure 2. Conference simulation: first paper session of the day. Participants have moved to one of six parallel sessions, according to interest.

The common-interest network was also used to determine a randomly generated plausible timetable – the actual timetable of any specific conference day was not used. Instead, 96 papers (out of a total of 445) were scheduled in six parallel streams of four sessions (early and late morning, early and late afternoon), each with four related papers. The first paper in each session was chosen randomly, and subsequent papers were chosen to be similar in topic (as determined by the common-interest network).

Paper sessions were interspersed with, and bookended by, coffee and meal breaks. During the breaks, participants converse on topics of common interest with other attendees with whom they have actual social network links. No plenary conference sessions were modelled. Figures 1 to 5 show snapshots of the simulation in progress.

2. TEXT PROCESSING

Since the electronic proceedings of the MODSIM 2005 conference has filenames based on the name of the primary author, a common-interest network of authors can easily be extracted, based on textual similarity of papers. Processing the documents and building the network was done by a Java program, using the PDFBox (2008) library for parsing PDF files.

For our purposes, the textual similarity of two papers was calculated using the correlation between *feature vectors* extracted from each paper. The feature vectors contained the relative frequency of word-pairs which:

- occurred at least 25 times in the whole proceedings (to rule out rare phrases);
- occurred in at least two papers (since words unique to one paper a useless for similarity calculation); and

- contained at most one "weak word" (defined to be numbers, words of less than three letters, and the common words "the," "and," "this," "are," "for," and "there").

Including rarer phrases gave excessively long feature vectors, slowing text processing, and giving poorer results. Scientific topics referred to in a paper generally involve key phrases which occur several times in the proceedings as a whole.

Work by Lee *et al.* (2005) suggests that using phrases of six or more words would be more effective than word pairs for assessing document similarity. However, in our work, longer phrases gave poorer results than word pairs. Longer phrases seem to be more indicative of similar writing styles or common origin of text than they are of topic similarity. Nevertheless, longer phrases are likely to be useful in our planned future work.

We converted the correlations between papers to an author common-interest network by thresholding correlations at a value of 0.0015. Only primary authors were considered, and where authors wrote multiple papers, the best-matching paper was used. Since the correlations act as link weights or values, the network is a valued network (Wasserman and Faust, 1994; Dekker, 2005a). For correlations involving the present author (Dekker, 2005b), the threshold of 0.0015 seemed to give the best balance between false negatives and false positives.

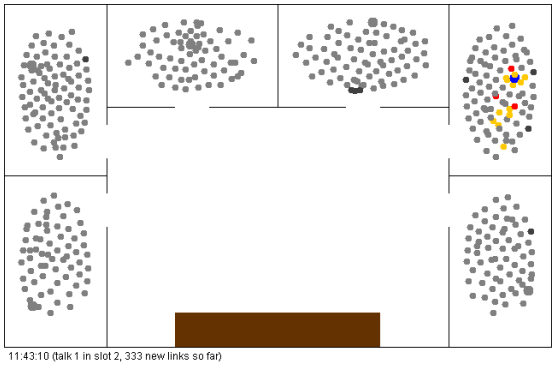

Figure 3. Conference simulation: the present author (blue) presents a paper. In this run, all participants with similar interests (according to the generated network) attend, and most switch from red to yellow as actual social linkages develop from discussion. Dark grey agents are readjusting their position within the seminar room.

The highest correlations observed involved authors who collaborated – that is, situations where one paper was written by Smith and Jones, and another by Jones alone. Correlations as high as 0.35 between Smith and Jones were observed in such cases, as a result of shared authorship.

Correlations involving the present author ranged from 0.0015 to 0.0071. Within this range were 15 authors. Ten of these indeed involved closely related papers, with presentations attended by the present author at MODSIM 2005. Four were unrelated work, where the correlation was the result of ambiguous terminology (such as varying meanings of the word "network"). The paper with the highest correlation (Aldridge, 2005) involved the shared topic of scale-free networks, but the present author became aware of the presentation only after it had been given.

The author common-interest network produced by text processing is clearly far from the "true" common-interest network, nor is it a perfect indication of which papers a conference participant will be strongly interested in. However, since participant behaviour involves random factors such as not being aware of a relevant presentation, the generated network is felt to be a sufficient approximation for our proof-of-concept simulation purposes. Comparison to the present author's diary notes indicates that the network is at least indicative of potential relationships between conference participants. It therefore suffices to produce at least partially realistic simulations, though not ones which allow conclusions to be drawn about individual participants.

3. SIMULATION DESIGN

The textual processing of the MODSIM 2005 conference proceedings provides an approximate common-interest network between participants. During the conference, participants with similar interests meet and exchange ideas, so that this potential social network gradually transforms into an actual social network. As with the text processing, the simulation design was guided by diary notes taken during the MODSIM 2005 conference. While these were adequate for a proof of concept, a more accurate simulation would require somewhat more detailed observations. At the time the diary notes were made, a conference simulation was not anticipated, and so considerable information was not recorded.

In developing the simulation, a number of simplifying assumptions were made. First, we assume that participants with similar interests make contact only during the question times following paper presentations. This is not entirely realistic, since many people, including the present author, will introduce themselves to total strangers during conference break periods. Furthermore, in real life introductions are also an important mechanism for making contacts: if Jones knows both Smith and Brown, she is likely to introduce Smith and Brown to each other when the opportunity arises. However, diary notes suggest that, at least for the present author, question periods are the primary contact mechanism.

Diary notes also indicate that for the present author, about 20% of participants with similar interests had been met prior to the conference, about 70% were met during the conference, and about 10% were not met, for various reasons. Based on this extremely approximate data, participants in the simulation were given, at random, a 20% chance of knowing each other prior to the conference, and where they did not, a 40% chance of making contact during the simulated day (the chance of not meeting during a four-day conference is thus $0.8(0.6)^4 = 0.1$).

During the paper presentations, participants attend the paper by the author to whom they are most strongly linked in the common-interest network. If there are no linked authors, they choose a paper at random. This leads to an unrealistically even distribution of participants across seminar rooms, which could perhaps have been avoided by biasing these random choices, for example by the overall popularity (number of links) of papers.

The simulation involves 448 agents, including 398 authors and 50 randomly chosen "students" of authors. A more accurate simulation would follow the actual conference demographics. Agent movement is calculated in 1-second timesteps, with the display updated every 10 seconds.

Agents have a variety of *goals*, dependent on the time within the simulated day. For example, during seminar sessions, the goal is to attend a selected paper, while during breaks the goal is to obtain refreshments. Agents with an unsatisfied goal move towards a region associated with satisfying that goal. Agents with unsatisfied goals are highlighted in dark grey by the simulation tool.

The simulation tool was written in Java, so that it could be provided as a Web applet as well as a standalone program.

4. SIMULATION RESULTS

Apart from demonstrating that textual data repositories can be used to build agent-based

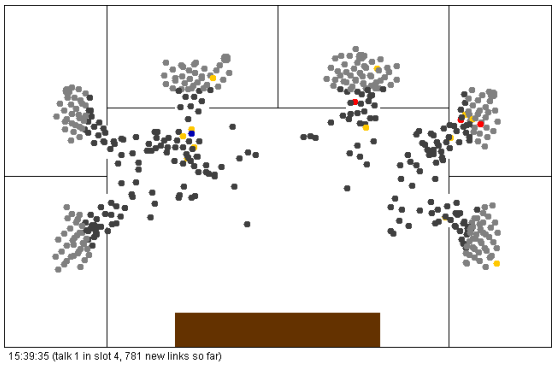

Figure 4. Conference simulation: participants leave afternoon coffee to attend the final session. Dark grey agents have not yet achieved their goal of entering their chosen session.

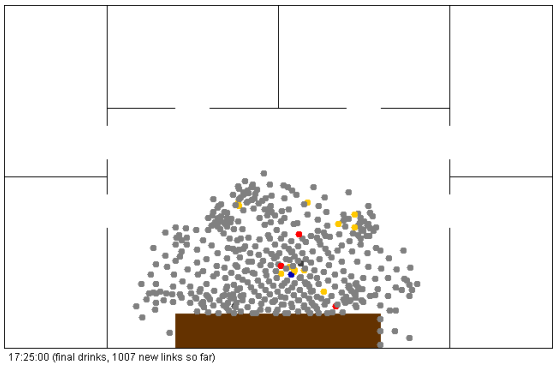

Figure 5. Conference simulation: drinks at the end of the day. With 1007 new social linkages developed during the day's discussions, participants cluster into conversational groups of people with similar interests (compare Figure 1).

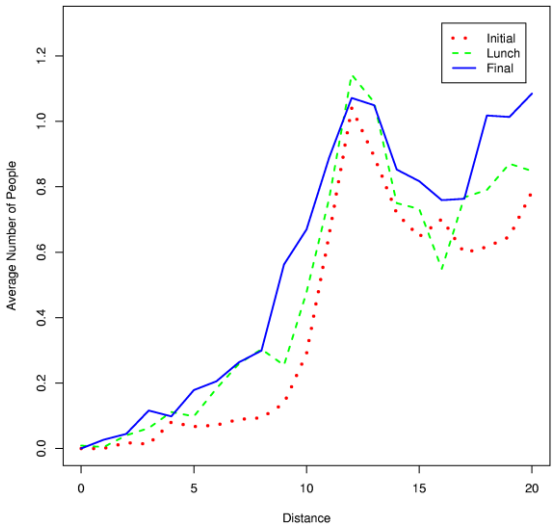

Figure 6. Conference simulation: variation in the distribution of people with distance for three break periods during the simulated day.

simulation models, the main question of interest was whether agents in the simulation would display realistic changes in behaviour from one break period to the next.

Qualitatively, the simulation displays behaviour characteristic of a real conference – changes in dynamics do indeed occur as participants get to know each other. Comparing Figure 1 and Figure 5 shows how the social linkages that develop cause conversational clusters of people to emerge. On some simulation runs, these conversational clusters impeded the flow of people to paper sessions, and caused delays in attendee arrival. Such phenomena are not unknown in real conferences, but in the absence of accurate observational data, the frequency of these phenomena cannot be compared against the real world.

Figure 6 highlights the changes in dynamics during break periods, for one simulation run. It shows the average number of people at a distance d from the average participant. The peak at $d = 12$ reflects the preferred interpersonal distance between strangers, programmed into the simulation. Figure 6 shows that during the course of the day, participants cluster more closely together. In particular, at lunch and at the end of the day, there are more people at a conversational distance of $d \approx 6$ than there were in the early morning. These increasing conversations with participants met during paper sessions are characteristic of a real conference, though again they cannot be assessed quantitatively without more accurate observational data.

5. TIMING DATA FROM TEXT DATA

For more general organisational simulations, textual data has the potential to identify timing information. This provides another important case where textual data can substitute for detailed observations. In particular, we are interested in the case where a network of agents produces textual documents, with each document being derived from previous documents produced by other agents, as part of a workflow process.

For each agent, we can associate a stream of text documents. Workflow linkages between agents are then revealed by correlations between the document streams. Furthermore, comparison of the document streams can provide an average *time lag* between the streams.

As the preliminary investigation of such analysis, we examined two RSS Internet news feeds collected on 25 September 2008, one from MSNBC, and the other from the Australian Broadcasting Corporation (ABC). Each feed contained 200 short news items. Because of the brevity of these "text documents," comparison was done by identifying common two-word phrases. News items were deemed to match if at least three common word pairs were found, with the number of common word pairs indicating the strength of the match. The best match involved this news item from MSNBC:

> 4:15 AM – "A young **Canadian man** was **found guilty** Thursday of knowingly **participating in a group that** was accused of plotting **to storm** Canada's **Parliament and behead the prime minister**."

The corresponding news item from the ABC was:

> 7:08 AM – "A **Canadian man** has been **found guilty** of **participating in a** terrorist **group that** allegedly planned **to storm Parliament and behead the Prime Minister**."

Among matching news items, the mean time lag between the MSNBC and ABC news feeds was 150 minutes. This is very similar to the time lag for the single news item quoted above, where the ABC's story followed 172 minutes after MSNBC. Naturally, this time lag of about 150 minutes does not demonstrate that the ABC bases its news items on MSNBC. In fact, they are likely to use common news sources.

Pairs of news items with time lags around 150 minutes had on average 0.2 matching word pairs, compared to 0.1 overall. This represents a rather weak correlation between the two news feeds.

Given a set of agents, each with a time-stamped series of output documents, we can infer a network of correlation/delay pairs, by thresholding at some minimum correlation, much as was done in Section 2 above. This network can be further simplified by considering triangular relationships.

Consider three agents A, B, and C, each with a time-stamped series of output documents. Given strong correlations A→B and B→C, with time lags t_1 and t_2 respectively, a weak correlation A→C with time lag approximately $t_1 + t_2$ can be disregarded, since it is a consequence of composing the two stronger correlations. Similarly, given strong correlations A→B and A→C with time lags t_1 and t_2, a weak correlation B→C with time lag approximately $t_2 - t_1$ can also be disregarded. The relationship between the MSNBC and ABC news feeds is likely to be an example of this kind of weak correlation.

As a result of simplifying such triangular relationships, a workflow network with time lags can be inferred, and this can form the basis of an agent-based simulation of the workflow process. A related inference process is described by Rowe *et al*. (2008).

6. DISCUSSION

We have demonstrated that analysis of textual data can be used to construct agent-based social simulation models. In particular, we have used the proceedings of the MODSIM 2005 conference to construct an agent-based simulation of one notional day of that conference. The conference simulation shows realistic emergent phenomena, such as increasing clustering of conversing participants during the course of the day.

The use of textual data does not completely eliminate the need for observational data, however. Although the use of text documents can significantly reduce the need for observation, some observational data is still needed. For the MODSIM 2005 conference simulation, limited diary notes on the actual conference were used, but more extensive notes would have assisted in producing a more detailed simulation. Nevertheless, we have demonstrated that the analysis of textual data can substantially reduce the amount of observational data required, and therefore significantly simplify model development.

For modelling workflow processes with an organisational simulation, time-delay information is required. A very simple analysis of two RSS Internet news feeds shows that such timing information can also, in principle, be inferred by textual data analysis. Textual analysis thus potentially provides a powerful approach to building organisational workflow models, and we intend to investigate this approach in further work.

REFERENCES

Aldridge, C. (2005), "Scale-Free Networks Using Local Information for Preferential Linking," in Zerger, A. and Argent, R.M., eds., *MODSIM 2005 International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand, December, pp 1196–1202, ISBN: 0-9758400-2-9, www.mssanz.org.au/modsim05/papers/aldridge.pdf

Dekker, A.H. (2005a), "Conceptual Distance in Social Network Analysis," *Journal of Social Structure*, Vol. 6, No. 3: www.cmu.edu/joss/content/articles/volume6/dekker/

Dekker, A.H. (2005b), "Network Topology and Military Performance," in Zerger, A. and Argent, R.M., eds., *MODSIM 2005 International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand, December, pp 2174–2180, ISBN: 0-9758400-2-9, www.mssanz.org.au/modsim05/papers/dekker.pdf

Lee, M.D., Pincombe, B.M., and Welsh, M.B. (2005), "An empirical evaluation of models of text document similarity," in Bara, B.G., Barsalou, L.W., and Bucciarelli, M. (eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pp. 1254–1259.

PDFBox (2008), PDFBox – Java PDF Library, version 0.7.3, www.pdfbox.org

Rowe, N.C., Sjoberg, E., and Adams, P. (2008), "Automatically Tracing Information Flow of Vulnerability and Cyber-Attack Information through Text Strings," *Proc. 13th International Command and Control Research and Technology Symposium*, www.dodccrp.org/events/13th_iccrts_2008/CD/html/papers/117.pdf

Wasserman, S. and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Zerger, A. and Argent, R.M., eds. (2005), *MODSIM 2005 International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand, December, ISBN: 0-9758400-2-9, www.mssanz.org.au/modsim05