Force-Based Clustering for Transitive Identity Mapping

H. Van Dyke Parunak
Soar Technology
3600 Green Court, Suite 600
Ann Arbor, MI 48105
+1 734 887 7643
van.parunak@soartech.com

Sven A. Brueckner
Soar Technology
3600 Green Court, Suite 600
+1 734 887 7642
sven.brueckner@soartech.com

ABSTRACT

In most information retrieval systems, software processes reason about passive data. Our approach instantiates each piece of information as an agent that actively seeks to organize itself with respect to other agents (including queries). Imitating the movement of bodies under physical forces, we describe a distributed algorithm ("force-based clustering," or FBC) for dynamically clustering and querying large, heterogeneous, dynamic collections of entities. The algorithm moves records in a virtual space in a way that estimates the transitive closure of the pairwise comparisons. We demonstrate FBC on a large, heterogeneous collection of records, each representing a person. We have some information about a person of interest, but no record in the collection directly matches this information. Application of FBC identifies a small subset of records that are good candidates for describing the person of interest, for further manual investigation and verification.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence]:** Distributed Artificial Intelligence – *Multiagent Systems*
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Retrieval models, Search process*

General Terms

Algorithms.

Keywords

Clustering; transitive mapping.

INTRODUCTION.—Many real-world search problems depend on inexact matches against heterogeneous data sources. Consider, for example, a person of interest (POI) who is traveling away from home. His cell phone died shortly before his departure, and he has borrowed a phone from a friend. During his travels, the POI leaves the number of the borrowed phone with someone whom he wants to contact him, but the number is partially illegible. The kind of information we might have includes a directory of cell phone numbers with the name and address of the registered user, regular telephone white pages with names, addresses, and landline phone numbers, airline reservation information with names, origin, destination, flight number, and time, and hotel registrations including name and credit card number. The names in these different sources are not represented in a consistent format, and there may be spelling errors. Identifying the POI requires computing the transitive closure of numerous relations. Here is one, but not by any means the only, route to a solution: the cell phone number leads to addresses for people who have some connection with the POI. These addresses may indicate the home area of the POI. Searching flight records for people who traveled from that area to the area where the phone number was left would generate a set of candidates for the POI, which might be further narrowed by matching against hotel registrations in hotels that are particularly near the location where the number was left.

Constructing and reasoning over such scenarios is combinatorially prohibitive, and too slow for emergencies (such as tracking the outbreak of an epidemic or disrupting a terrorist attack) where it is critical to find the POI quickly. Our subsymbolic approach does not require such complex preparation. We instantiate each entity as a software agent in an abstract low-dimensional space (a three-dimensional torus wrapped in four dimensions). The agents compute virtual "forces" among themselves, and move in response to those forces. The transitivity of these forces brings together agents whose similarity may not be documented directly, but that are linked by a chain of similar agents.

PROBLEM STATEMENT.—Imagine the following scenario:

> An unidentified male visited a medical clinic and signed in with an illegible signature and partially illegible phone number. Before receiving attention, the individual exited the facility. Later, another patient showed symptoms of an influenza-like illness consistent with a potentially deadly and highly contagious virus. In reviewing the sign-in log, staff discovered an entry which was unaccounted for, who apparently introduced the infection to the clinic.

For the good both of the mystery patient and of the general public, it is essential to identify this person as quickly as possible.

An anonymous sponsor provided us with eight databases (DBs), containing varying combinations of name, address, phone number, and DB-specific record identifiers (Doc-IDs) for fictitious individuals, but concealing the identity of the POI. The total number of records is on the order of 350,000.

Cursory examination of the data indicates that some DB-specific keys are shared both within and across databases, and some names appear to be variant spellings (e.g., "Tom F. Tuk" and "Tolman Fredegar Took" share other information). Only the last two digits of the phone number are illegible, but 104 records have phone numbers that could match the available number, some associated with different names or addresses, suggesting errors in the data. The phone numbers are all in a different state than the clinic.

Our task is to develop a prioritized list of people with contact information whom authorities should contact in an effort to reach the mystery patient as quickly as possible.

TECHNICAL APPROACH.—Our approach is motivated by physical forces. Interacting physical entities show several characteristics, including *repulsion* of entities that are very close together, *decrease* of interaction with distance, and *integration* of multiple forces. Each record in our database becomes an agent. We distribute them in an abstract space, define virtual forces among

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013),* Ito, Jonker, Gini, and Shehory (eds.), May 6–10, 2013, Saint Paul, Minnesota, USA.

them, and let them move. Similar records will move closer to one another, pulling their neighbors with them (and thus providing transitive closure). We query the system by inserting a query record that contains what we know about the POI, letting the agents move until the system has converged, and retrieving records that end up close to the query. The closer a record is to the query, the higher we rank it in our list of persons to contact.

We emulate the features of physical movement.

- Extremely close agents repel one another, keeping similar records from collapsing to the same location.
- Interaction decreases with distance, so most interactions are local. Local interactions reduce the set of agents with which a given agent effectively interacts, allowing their influence to be felt in fewer steps.
- The concept of multiple forces lets us handle heterogeneous records with varying field contents. Integration of these forces through agent movement allows transitive interactions among records whose fields do not directly overlap.

Our implementation, discussed in detail in the full paper, includes similarity computation, force definition, distributed execution, and convergence detection. The full paper also discusses the relation between FBC and other technologies, including semantic analysis, cluster analysis, and multi-dimensional scaling.

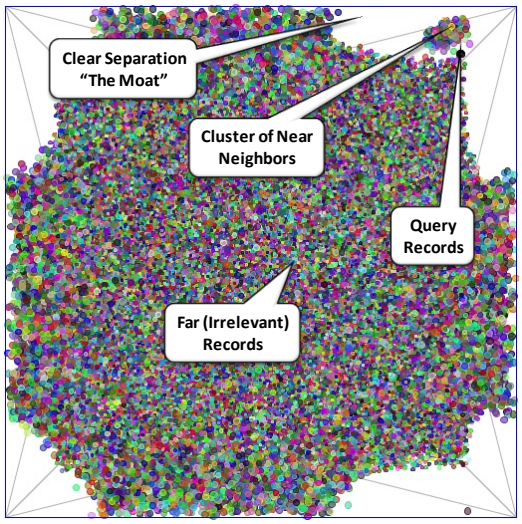

Figure 2: Converged Location of all 350k Records in Clustering Space

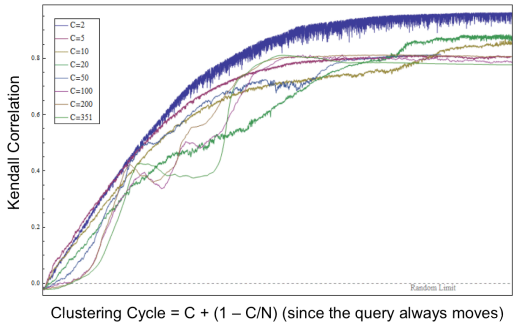

Figure 1: Kendall Correlation over Clustering Cycle * *C*.

EVALUATION.—FBC is inherently parallel and can be distributed for essentially linear (with the number of processors) performance gains over large scale networks and potentially deployed into a MapReduce/Hadoop cloud-computing environment. Our experiment used three standard WinTel PCs to execute 4 clustering processes each and an additional PC to run the MySQL database with the 350k records and their clustering coordinates. In this small setup, we arrived at the results reported here in less than two days execution even though one PC (4 processes) failed due to network problems after less than 8 hours. The clustering space is a unit (1x1x1) box with all 6 faces wrapped. Figure 2 shows the raw result. It highlights the common location of the three query records (phone, address, phone+address) in the upper right corner. Adjacent to the queries, the information matching process highlights a relatively small set of nearby neighbors (data records) as relevant, clearly separated by a "Moat" from the rest of the data. Analysis of this result in the full paper shows that

- The process is highly selective, discarding records that are superficially similar;
- The mass of irrelevant records does contain structure that can support parallel queries;
- The cluster of near neighbors includes records with no direct similarity, supporting transitive closure of the process;
- We found the person of interest, as later confirmed by the sponsor who provided the data.

We assessed the convergence of FBC with an artificial data set of 350 color (RGB) data records, groups into seven clusters, and one query record. We start the experiment with a random arrangement of the records' agents in cluster space and run to (manually determined) convergence, for various values of the parameter *C* that determines how many randomly selected agents are allowed to interact in each cycle. We assess the quality of clustering using Kendall correlation. Figure 1 shows the exponential shape of convergence. It also assesses the impact of parallelization by scaling the x-axis for each data series by a factor of *C* and correcting for the movement of the query record. Thus scaled, the convergence curves trend very close to each other, suggesting a nearly linear speed-up with the number of processors.

CONCLUSIONS.—The current project demonstrates the ability of FBC to find transitively related groups of records in a distributed environment that can scale to handle massive data. The full paper discusses opportunities for extension, including more disciplined *weighting* of different similarity components, provision for further *human interaction*, *distributing data* as well as processing, and running with *dynamically changing data*.

Many important applications in epidemiology and domestic security require the ability to discover transitive linkages across heterogeneous databases rapidly, without reasoning explicitly about possible scenarios. Instead of reasoning about the various records, Force-Based Clustering (FBC) turns each record into a software agent that moves in an abstract information space in response to the net "force" it feels from other agents. These forces in turn are defined by generic similarity measures over commonly occurring fields, measures that can readily be defined in advance and applied quickly to available information. The agent interactions can be distributed over many processors to speed the clustering process. Application of this approach to a synthetic data set (provided by an anonymous sponsor external to our research group) allows us to identify the person of interest.

The full paper is available at
http://abcresearch.org/papers/AAMAS2013TIM.pdf.