Guessing at the Content of a Million Books

Patrick Juola

Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, Pennsylvania, USA
juola@mathcs.duq.edu

Abstract. The recent growth in digital scholarship has made literally millions of books available to readers. But the implications of this, paradoxically, are that reading becomes more difficult. No human can possibly read and understand a million books. This is particularly problematic in literary scholarship, where "reading" a text requires much more than simple content extraction, but may require identifying and explaining patterns of thought and expression across many different works.

We propose a new computer-mediated form of reading, based on automatic pattern extraction. A recent example of this is the "Adam" robot (BBC, 2 April 2009). Other examples include Eurisko[1] and Graffiti[2] to perform automatic research in mathematics.

The Graffiti program, in particular, researches graph theory through the generation and testing of conjectures. The program creates random, template-based conjectures, which are then tested against a large collection of graphs. Any conjectures that survive this set of tests are published. Graffiti, it should note, does not prove any conjectures, but will provide a list of statements that appear to be true; mathematicians are encouraged to prove or disprove them. Since its inception, Graffiti has listed over 1000 different conjectures and inspired more than 100 published papers.

A similar paradigm allows us to conjecture the existence of patterns in writing. We know, for example, that language varies over time, over genre, and over authorial gender[3] in many specific ways. But Roget's thesaurus lists more than 1000 different semantic "categories", most of which have never been studied in the context of gender and language. For example, we are aware of no study of the use of animal terms (Roget category III.iii.1.2/366). Do men and women's speech differ in this regard? Having constructed this conjecture, it is easy for a computer to test this. If true, this is an interesting finding in need of explanation.

A prototype system to do this initial research, the Conjecturator[4], has been constructed; some sample conjectures are available at http://www.twitter.com/conjecturator. Any or all of these published conjectures could serve as the basis for an interesting explanatory paper. We offer this as an example of a new paradigm in reading and scholarship; an opportunity to separate rote reading (which can be done by computer) from the actual scholarly and intellectual work.

Keywords: reading, *Graffiti*, *Conjecturator*, prototyping, computer-mediated reading, application, text analysis, textual patterns and algorithms.

The recent growth in digital scholarship has made literally millions of books available to readers. But the implications of this, paradoxically, are that reading becomes more difficult. No human can possibly read and understand a million

books. We can make the usual hyperbole-filled calculations – i.e, a person reading ten books a week, fifty weeks a year, would take exactly two thousand years to read a million books. This is particularly problematic in literary scholarship, where "reading" a text requires much more than simple content extraction, and may require identifying and explaining patterns of thought and expression across many different works. Of course, no one seriously proposes to do this, but what is needed is a method of accessing the patterns (and underlying data) across millions of books without resorting to close reading.

One method that has been proposed is of course search technology; technology such as Google Books will make it possible for lexicographers to build a concordance of all uses of a particular word or phrase in the database.

More abstract queries such as "all uses of personification" are more problematic, but not unduly so -- if personification can be defined clearly enough (McCarty, 2003), computers can search for and find it.

What, however, of the creatively serendipitous discovery? A reader finds a passage or a pattern that sparks a train of thought, inspiring her to read and re-read other documents to refine, refute, confirm and explain her new idea. In fact, most of the interesting parts of scholarship are not in the simple observation, but in the refinement and explanation; for example, knowing that women use more tag questions (phrases like "isn't it?") and intensifiers (words like "very" or "really" or "extremely") (Glass, 1993) isn't as interesting as knowing why these differences arise. This provides an opportunity to separate rote reading (which can be done by a computer) from actual scholarship. Rote reading and observation can lead to possible new avenues for scholars to explore and potential new insights.

We propose a new computer-mediated form of reading based on automatic rote reading and pattern extraction. A recent example of this is the "Adam" robot (BBC, 2 April 2009). Other examples include Eurisko (Lenat, 1983) and Graffiti (Fajtlowicz, 1988), which perform automatic research in mathematics. As the title somewhat glibly suggests, we propose to reverse the ordinary order of pattern recognition by guessing at the existence of a particular pattern, then looking for evidence to support or refute this guess.

The Graffiti program, in particular, researches graph theory through the generation and testing of conjectures. The program creates random, template-based conjectures, which are then tested against a large collection of graphs. Any conjectures that survive this set of tests are published. Graffiti, it should note, does not prove any conjectures, but will provide a list of statements that appear to be true; mathematicians are encouraged to prove or disprove them. Since inception, Graffiti has listed over 1000 different conjectures and inspired more than 100 published papers.

A similar paradigm allows us to conjecture the existence of patterns in writing. We know, for example, that language varies over time, over genre, and over authorial gender in many specific ways. We do not, however, have a complete catalog of variation. Roget's thesaurus lists more than 1000 different semantic "categories", most of which have never been studied in the context of gender and language. For example, we are aware of no study of the use of animal terms (Roget category III.iii.1.2/366). Do men and women's speech differ in this regard? Having constructed this conjecture, it is easy for a computer to test this. If true, this is an interesting finding in need of explanation. If false, we've wasted nothing but the computer time necessary to disprove the conjecture.

We thus see that we can separate the process of conjecture generation (guessing about things that might be true in the corpus) from analysis and explanation. With this separation, we are presented with a new method of knowledge generation, as follows:

Phase 1: the computer guesses at the existence of a particular pattern across a large text corpus. This is easiest to do using a template (following Graffiti) such as “Property X appears more often in document type Y than in document type Z”, where X, Y, and Z are randomly generated.

Phase 2: the computer analyzes the corpus to see whether the conjecture is supported by analyzing the corpus of interest. Because this analysis can take place at computer speeds, a directed reading of the entire million-book corpus can take minutes or hours instead of centuries. Of course, most conjectures will turn out to be wild-eyed speculation with no support for them, and can be discarded immediately. But for those that are true,

Phase 3: the computer publishes the conjecture (possibly together with supporting evidence) for human scholars. Human scholars will decide, as with Graffiti, if the conjecture -- now presumptively promoted to “fact” -- is worth examining in detail and explaining.

A prototype system to do this type of reading, termed the Conjecturator (Juola and Bernola, 2009), has been constructed; some sample conjectures are presented here:

The word group cohabitation appears less in regional fiction novels than in early Victorian novels (9.406051929154446E-4)

The word group wrangle appears less in English female authored novels than in sensation novels (0.003028742030361964)

The word group perfumer appears less in English female authored novels than in English male author novels (0.019532152290108074)

The word group gunsel appears more in mid-Victorian novels than in bourgeois fiction (0.9998934112294822)

The word group happy hour appears more in realist novels than in bourgeois fiction (0.9999998397796231)

The word group steeled appears less in novels with Protestant issues than in feminist novels (0.01781726502881331)

The word group atheist appears more in English female authored novels than in psychological realism novels (0.9997023579705397)

The word group pennant appears less in bourgeois fiction than in satirical novels (0.009731536472060709)

The word group loutish appears more in novels with Protestant issues than in sensation novels (0.9899785414211395)

The word group impose appears less in American male authored novels than in bourgeois fiction (0.0038512615378674675)

The word group experience appears more in realist novels than in regional fiction novels (0.9803429183312695)

In each case, the number following the conjecture is the observed p-value of a statistical test establishing likelihood of the observed difference among the novel groups studied.

How were these numbers obtained? For these experiments, we used a corpus of Victorian novels and categorized them along “standard” divisions of authorship, genre, and time. (For example, "Jane Eyre" is a female-authored novel, a governess novel, a feminist novel, a domestic realism novel, an English-authored novel, and so forth.) Using a standard machine-readable thesaurus with about 30,000 categories, the computer tabulated the frequency with which the words representing a specific (random) concept (such as "atheist") in each document of that type. Simple statistics gives us a mean (average) and variance (deviation). Most conjectures show that the conceptual variation appears to be random or nothing more than chance predicts, but for some concepts there is a significant and as yet unexplained difference.

More are available at http://www.twitter.com/conjecturator. Any or all of these published conjectures could serve as the basis for an interesting

explanatory paper. We offer this as an example of a new paradigm in reading and scholarship; an opportunity to separate rote reading (which can be done by computer) from the actual scholarly and intellectual work.

Works Cited

Fajtlowicz, Siemion. "On conjectures of Graffiti." *Discrete Mathematics* 72 (1988). Print.

Glass, Lillian. *He Says, She Says : Closing the Communications Gap Between the Sexes*. Perigee Trade, 1993. Print.

Juola, Patrick and Ashley Bernola. "Conjecture Generation in the Digital Humanities." *Proceedings of DH*. College Park, MD. 2009. Print.

Lenat, Douglas. "EURISKO: A program that learns new heuristics and domain concepts." *Artificial Intelligence* 21 (1983). Print.

McCarty, Willard. "Depth, Markup, and Modelling." *Computers and the Humanities Working Papers* A.25 (2003). Print; rpt. *TEXT Technology* 12.1 (2003). Print.