2005 Paper 7 Question 12

Information Retrieval

The SNOWBALL algorithm uses bootstrapping from known tuples of named entities which stand in a well-defined relationship, in order to detect new tuples.

(*a*) Describe SNOWBALL's algorithm in detail, including the thresholds used in the single steps of the algorithm. [7 marks]

(*b*) The table below contains corpus examples of co-occurrences of organisation names (o) and location names (l). Consider a situation where SNOWBALL is applied to the corpus examples given here, when the only known tuples are <Microsoft, Redmond> and <Exxon, Irving>.

A	<l>Seattle</l>-based company <o>Boeing</o> offered ...
B	Yesterday, at <o>Microsoft</o>'s headquarters in <l>Redmond</l>, the deal was brokered ...
C	Though they had never been at <l>Redmond</l>, <o>Microsoft</o> showed them ...
D	In <l>New York</l>, <o>Microsoft</o> stock nosedived ...
E	When we arrived in <l>London</l>, <o>Exxon</o>petrol stations were ...
F	... met at <o>Microsoft</o> headquarters. In <l>Redmond</l>, ...
G	<o>Boeing</o>, <l>Seattle</l>, had no choice but to ...
H	In <l>New York</l>, <o>Intel</o> stock recovered ...
I	... due to arrive in <l>Irving</l>, <o>Exxon</o> executives might ...
J	<o>Boeing</o> headquarters in <l>Seattle</l> are air-conditioned ...
K	<o>Microsoft</o>, <l>Redmond</l>, made a statement ...
L	<o>Boeing</o>, <l>Seattle</l>, confirmed ...
M	<o>Microsoft<o>, <l>Redmond</l>, readily agreed ...
N	... <o>Exxon</o>. Although they had never in their whole life been in <l>Irving</l>, they ...
O	<o>Exxon</o>, <l>New York</l>, was a winner in our recent ...

Discuss which patterns get hypothesised and which new tuples this produces in the next iteration. Assume sensible thresholds. [6 marks]

(*c*) What happens to the result in part (*b*) if the sentence "*Microsoft's previous headquarters in Cincinnati were insured for 20 million dollars.*" gets added to the corpus? [3 marks]

(*d*) The SNOWBALL algorithm is to be applied to find tuples of person names and their professional positions from a large newspaper corpus. Would you expect SNOWBALL to work well on this task, and why? [4 marks]