

Relationen mellan IE och 'text mining'

Anette Hulth
hulth@dsv.su.se

26 juni 2000

1 Introduktion och sammanfattning i ett

Detta är en sammanfattning av det sista seminariet om informationsextraktion, vilket hölls den 8 juni 2000. Temat som behandlades var "relationen mellan informationsextraktion och text mining". Nedan kommer jag inledningsvis att ge några olika definitioner av *text mining*, så som beskrivet i litteraturen. Därefter kommer jag att redogöra för vilken eventuell plats informationsextraktion har i de olika definitionerna. I den avslutande diskussionen kommer jag att hävda att det är hart när omöjligt att fastställa relationen de båda ämnena emellan, eftersom begreppet *text mining* så som det används idag är alltför vagt samt eftersom de som anser sig syssla med detta område har sina egna definitioner av begreppet.

2 Vad är text mining?

För att utröna huruvida det finns en relation mellan *informationsextraktion* och *text mining* och hur då denna ser ut, måste först begreppet *text mining (TM)* redas ut. Begreppet *informationsextraktion (IE)*, som utförligt har behandlats under kursen gång, kommer i princip inte att förklaras i detta dokument. Det kan dock kort sammanfattas så som varandes: att fylla i fördefinierade mallar (eller *templater*) med relevant information. Detta förutsätter förstås att vi vet vilken typ av information det är vi söker.

För att reda ut begreppet *text mining* har jag tittat på fyra olika definitioner av detta område. Tre av dessa är skapade av vad man skulle kunna kalla för auktoriteter inom TM: Ronen Feldman; Marti Hearst; samt Yves Kodratoff. Den fjärde är hämtad från en australiensk student vid namn Mark Dixon. Vad man inledningsvis kan säga om dessa olika källor är att ämnet behandlas av var och en som om idéerna vore helt nya och framför allt att försök till definitioner tidigare inte har gjorts. Detta märks bland annat genom att de använder sig av olika namn på ämnet (*text mining* – Feldman; *text data mining* – Hearst; *knowledge discovery in texts* – Kodratoff; *document mining* – Dixon). Att de syftar på samma sak framgår genom att de på ett sätt eller annat alla nämner *text mining* som ett alternativt namn.

Hearst (1999) och Kodratoff (1999) diskuterar också det faktum att begreppet *text min-*

ing idag ofta används på ett något felaktigt sätt av kommersiella företag, där den aktivitet man åsyftar redan har ett eget namn, men att man trots detta gärna pratar om det så som varandes text mining, på grund av den positiva klang som begreppet har fått av “marknaden”. (Kodratoff (1999) nämner till exempel informationsåtkomst (*information retrieval*) som ett dylikt område.)

2.1 Text Data Mining

Enligt Hearst (1999) anser många att text data mining är data mining gjord på text. Men så är inte fallet, fortsätter hon. Data mining handlar om att hitta mönster i stora mängder data; att kalla detta för mining (eller brytning som det bör heta på svenska) är egentligen en ganska dålig metafor. När det däremot gäller text, så vill Hearst att mining-metaforen ska tolkas på ett mer bokstavligt vis: att bland en stor mängd gråsten hitta små guldklimpar i form av okänd kunskap.

Hearst betonar att kunskapen för att kunna kallas “tidigare okänd” inte får förekomma i något av de dokument man för tillfället arbetar med—då skulle ju författaren till detta dokument känna till detta sedan tidigare. Kunskapen måste således vara ny. Något som enligt Hearst (1999) kännetecknar text data mining är också att kunskapen säger något om verkligheten utanför dokumenten.

I figur 1 visas en översikt gjord av Hearst (2000) på vilka discipliner som ägnar sig åt icke-textuell och textuell data, respektive mönster och “guldklimpar”.

	Patterns	Non-novel Nuggets	Novel Nuggets
Non-textual Data	Standard data mining	Database queries	AI Discovery Systems
Textual Data	Computational linguistics	Information retrieval	Real text data mining

Figur 1: Vem gör vad? Från Hearst (2000).

Som ett exempel på hur text data mining skulle kunna se ut, beskriver Hearst (1999; 2000) forskning av en person vid namn Swanson. Genom att manuellt undersöka och länka samman information från en stor mängd rubriker i biomedicinska tidskrifter, har Swanson lyckats komma fram till att magnesiumbrist kan vara en orsak till migrän. Då Swanson lade fram denna teori var detta helt okänt, men senare medicinska försök har visat att så faktiskt är fallet. Hearst menar att genom att automatisera åtminstone delar av denna process så skulle vi få det som hon vill kalla text data mining.

2.2 Knowledge Discovery in Texts

Kodratoff (1999) kallar området för *Knowledge Discovery in Texts (KDT)*, analogt med Knowledge Discovery in Databases (KDD), fast på text. Till skillnad från Hearst så söker inte Kodratoff särskilja “vanlig” data mining och dito gjord på texter, utan anser att det i princip är samma sak som görs och att processen är densamma.

För att något i detta fall skall kallas kunskap (*knowledge*) menar Kodratoff (1999) att

“The knowledge extracted has to be grounded in the real world and will modify the behaviour of a human or mechanical agent”.

Kunskapen skall också vara möjlig att förstå (*understandable*) och kunna användas direkt (*directly usable*). Att man kan kalla det för en upptäckt (*discovery*) är för att den grundläggande metodiken är induktion. Kodratoff betonar också pluralformen på begreppet *texter* (*texts*). En förutsättning för att prata om KDT är alltså att vi har en större mängd texter över vilka vi letar kunskap.

Kodratoff (1999) ger ett exempel på hur KDT kan se ut. Verktuget *Tropes* har analyserat texter från ett par årgångar av *Le Monde*. Resultat (i form av regler) visar att tidningen inte använder instanser av begreppet *katastrof* (till exempel *översvämning* eller *olycka*) om man talar om Nordamerika, familjer, kvinnor, eller ekonomi. Däremot talar man om detta när det gäller cirka 300 andra koncept. Vad detta kan bero på låter Kodratoff dock vara osagt.

2.3 Text Mining

Feldman (1999a; 1999b) är den som har den vidaste definitionen av text mining, och dessutom den enda som konsekvent kallar området för just text mining. Han menar att text mining har utvecklats (eller snarare håller på att utvecklas) för att hjälpa oss att få ordning på all information i textform i dagens samhälle. Detta är kanske den största skillnaden mellan Feldmans definition gentemot de av Hearst (1999) och Kodratoff (1999). De två förstnämnda vill nämligen hitta eller skapa ny information, medan Feldman hellre vill lösa problemen med dagens informationsöverflöd.

Enligt Feldman (1999b) så bygger text mining på tekniker från data mining, maskininlärning, informationsåtkomst, naturligt språkförståelse, case-based reasoning, statistik och knowledge management. Processen kan grovt delas upp i fyra steg: förbehandling (textkategorisering eller termextraktion); lagring och indexering; analys (kan genomföras med en mängd olika tekniker); samt visualisering.

För det symposium i text mining under *IJCAI'99*, för vilket Feldman delvis var ansvarig¹, trycktes 22 artiklar eller korta artiklar. Av dessa är det i majoriteten svårt att se var själva “text miningen” ligger. Utan tvekan behandlas i samtliga artiklar något steg i processen—till exempel visualisering—men så mycket mining talas det i flertalet inte om.

2.4 Document Mining

Dixon (1997) menar att *document mining* letar efter mönster och tidigare okänd kunskap i ostrukturerade texter. Som exempel på vilka frågor man vill kunna besvara ger han: “Hur många terroristattacker begicks under 1995?” eller “Gör ett företag en bättre förtjänst genom att byta chef ofta?”. Terroristattacker och positionsbyten är typiska domäner för informationsextraktion så som den har bedrivits i konferensserien MUC (*Message Understanding Conference*).

¹I programkommittén för detta symposium fanns även Marti Hearst och Yves Kodratoff

Document mining kombinerar tekniker från: informationsextraktion; informationsåtkomst (IR); naturligt språkprocessande; och textsammanfattning. Mining-processen är stegvis: IR – hitta för uppgiften relevanta dokument; IE – extrahera information från dessa (med hjälp av templatser); mining-steget – hitta mönster i de fyllda templaterna; samt slutligen att tolka det funna mönstren, och på så sätt omvandla det till kunskap.

3 Hur förhåller sig informationsextraktion till dessa definitioner?

Hearst (2000) menar att informationsextraktion möjligtvis kan vara användbart för text data mining, men i huvudsak går IE ut på att fylla i fördefinierade mallar. Som tidigare diskuterats så betonar Hearst det okända i kunskapen, och hon menar att det hon förespråkar snarare är mallar vilka man inte från början kan veta vilka de är—alltså snarare okända eller oförutsedda mallar. Hearst (2000) anser alltså inte att informationsextraktion är en förutsättning för eller ett nödvändigt steg i processen. Hon kan möjligen även mena att IE kan verka begränsande, eftersom vi vill hitta okänd kunskap och IE, som sagt, förutsätter att vi vet vad vi vill hitta.

Kodratoff (1999) menar att informationsextraktion kan utgöra en del av ett generellt text mining-system. Han påpekar också att IE i sig är ett icke-trivialt problem som redan har sin rättmätiga plats i NLP-samfundet.

Feldman nämner informationsextraktion endast i sin inbjudan till det tidigare nämnda symposiet på IJCAI'99 (Feldman, 1999b), då ett av de tema som han föreslår är: *”Use IE to better capture the major themes of the documents”*. Det saknas dock bidrag av denna karaktär i artikelsamlingen. Det är också möjligt att Feldman tänker sig att man kan använda IE-tekniker för det steg som han kallar förbehandling.

Dixon (1997) är den enda som explicit nämner informationsextraktion i sin definition av området, där han alltså ser IE som ett nödvändigt och självklart steg i mining-processen.

4 Diskussion

Text mining har under de senaste två-tre åren av många blivit utsett till det som ska rädda oss från det informationsöverflöd som hotar att förgöra oss. Som med många beteckningar som fått en “hype-stämpel”, används även begreppet *text mining* för en mängd olika aktiviteter vilka inte tillnärmelsevis liknar något som skulle kunna kallas text mining. Problemet blir inte mindre då det saknas en överenskommen och fastslagen definition, till vilken man kan hänvisa.

Något som utförligt diskuterades under seminariet var huruvida det man vill göra med texter är något annat än det som data mining söker göra med de datamängder som det arbetar med. Till exempel diskuterades hur pass mycket processerna skiljer sig åt, beroende på om data är i textform eller i form av numeriska värden. Syftet med denna diskussion var att komma till insikt om hur motiverat det är att hitta på ett nytt namn på området, och hur detta då detta område skulle kunna definieras.

Vad gäller informationsextraktionens roll i text mining, fick detta problem stå tillbaka lite för den just beskrivna diskussionen. Att IE kan fungera som en användbar komponent i processen torde vara odiskutabelt. Det är dock tveksamt att det är en nödvändig komponent. Kanske är det dock så att först när vi har en definition som alla kan enas kring, och i och med detta har ett klart definierat mål, kan vi i mer generella ordalag svara på i vilken utsträckning informationsextraktionstekniker kan vara till nytta, eller till och med en förutsättning, för en lyckad text mining-process.

Referenser

Feldman, R. 1999a. Practical Text Mining. Tutorial at EACL'99.

Feldman, R., editor. 1999b. *Text Mining: Foundations, Techniques and Applications*. Workshop at IJCAI'99.

Dixon, M. 1997. Document Mining Technology. URL:
<http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/main.html>. Hämtad 1999.

Hearst, M. 2000. Text Mining Tools: Instruments for Scientific Discovery. I *IMA Text Mining Workshop*. Power-pointpresentation.

Hearst, M. 1999. Untangling Text Data Mining. I *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland.

Kodratoff, Y. 1999. An Application to Knowledge Discovery in Texts. Lecture at ACAI'99.