Executive summary

Entity Profiling for Intelligence Using the Graphical Overview of Social and Semantic Interactions of People (GOSSIP) Software Tool:

Peter Kwantes; Phil Terhaar; DRDC Toronto TR 2010-188; Defence R&D Canada – Toronto; November 2010.

Introduction or background: The reports of interest to the intelligence analyst or Influence Activities analysts will generally be analysed to uncover the people and organisations discussed in the report collection, and how these entities are connected to one another. While such information is useful, it does not reveal information about the entities that can be uncovered through automated methods. In this report, we'll demonstrate that an unsupervised model of semantic memory can be used to generate profiles of entities discussed in document collections. Semantic memory refers to memory for the things one knows as opposed to memory for the things one can remember. The past 20 years have seen great advances in our understanding of semantic memory. So much so, that the latest models can create semantic representation for terms in a completely unsupervised fashion. That is, the models can figure out what terms are semantically similar without the model builder hand-wiring any associations among them. GOSSIP is a software tool developed at Defence Research and Development Canada (DRDC) - Toronto that allows the user to see the connections that exist among entities discussed in a large collection of documents. GOSSIP has a model of semantics working in the background processing the documents in the collection. Over the thousands of documents of a collection that are processed, it forms semantic representations for terms and entity names. The semantic representations form a basis upon which to filter documents or entities. In this report, we show that GOSSIP's semantic representations of entities and documents discussed in the text can be queried to find out what concepts connect entities, and more interestingly, it can generate profiles across a set of user-defined qualities. We tested GOSSIP by conducting two empirical studies in which subjects were asked to make judgments about famous names, and about the concepts that connect pairs of famous names.

Results: We found that the information GOSSIP extracted from the document collection was, for the most part, in line with the domain knowledge provided by subjects. In other words, humans and GOSSIP were in close agreement about the material discussed in the document collection.

Significance: We take the results reported here as clear evidence that GOSSIP is a potentially useful tool for quickly establishing situational awareness about people, places and groups discussed in large document collections (situation reports or open source media). Using GOSSIP will help analysts in the Canadian Forces to gain situation awareness about a domain in a timely manner without sacrificing accuracy.

Future plans: It is our goal that the capabilities embodied by GOSSIP will at some point be integrated as a service in existing analysis tools being developed for, and exploited by, the Canadian Forces.

Sommaire

Entity Profiling for Intelligence Using the Graphical Overview of Social and Semantic Interactions of People (GOSSIP) Software Tool:

Peter Kwantes; Phil Terhaar; DRDC Toronto TR 2010-188; R & D pour la défense Canada – Toronto; Novembre 2010.

Introduction ou contexte : En règle générale, l'analyste du renseignement ou celui des activités d'influence examine le corpus de rapports qui l'intéressent dans le but d'y repérer le nom des personnes et des organisations qui y sont mentionnées et de découvrir des « connexions », c'est-à-dire les liens qu'elles présentent entre elles. Bien qu'une telle information soit utile, elle n'apporte rien de plus que ce que des méthodes automatisées permettent d'apprendre à leur sujet. Le présent rapport nous permet de montrer qu'un modèle de mémoire sémantique appliqué sans supervision peut servir à générer les profils des entités mentionnées dans un corpus de documents. Par mémoire sémantique, nous entendons la mémoire qui procède de la connaissance, c'est-à-dire ce que l'on sait, par opposition à celle qui relève des souvenirs, c'est-à-dire ce dont on se souvient. Au cours deux dernières décennies, le savoir humain a réalisé des progrès remarquables dans le domaine de la mémoire sémantique, à telle enseigne que les modèles les plus récents sont en mesure de créer une représentation sémantique des termes en l'absence de toute forme de supervision. Autrement dit, ces modèles découvrent eux-mêmes les termes qui présentent des similitudes sémantiques sans que le concepteur ait à introduire des associations entre ces termes. Mis au point à RDDC Toronto, GOSSIP est un logiciel qui permet de visualiser les connexions entre les entités mentionnées dans un corpus comptant un grand nombre de documents. Son fonctionnement repose sur un modèle sémantique exécuté en arrière-plan qui traite les documents du corpus. Le logiciel crée une représentation sémantique de chaque terme et nom d'entité que contiennent les milliers de documents ainsi traités. De telles représentations sémantiques forment la base à partir de laquelle sont filtrés les documents et entités. Dans le présent rapport, nous montrons que GOSSIP permet d'interroger les représentations graphiques des entités et des documents mentionnés dans un texte et de découvrir ainsi les notions qui lient les entités entre elles. Plus intéressant encore, il permet de générer des profils répondant à un ensemble de caractéristiques préétablies. Nous avons vérifié l'efficacité de GOSSIP sur ces deux plans en menant autant d'études empiriques aux cours desquelles les personnes interrogées devaient exprimer leur opinion sur diverses célébrités et sur la nature des liens qu'évoquent dans leur esprit divers noms de célébrités appariés.

Résultats : Nous avons constaté que les renseignements que GOSSIP extrait du corpus de documents correspondaient dans une large mesure aux connaissances des personnes interrogées dans le domaine. En d'autres mots, ces personnes et GOSSIP étaient presque unanimes quant au contenu mentionné dans le corpus de documents.

Portée : Nous sommes d'avis que les résultats présentés ici démontrent clairement que GOSSIP peut s'avérer utile pour élaborer rapidement une connaissance de la situation sur des gens, des endroits et des groupes mentionnés dans un corpus volumineux de documents, qu'il s'agisse de comptes rendus de situation ou de contenu provenant de sources ouvertes (médias). Grâce à

GOSSIP, les analystes des Forces canadiennes pourront acquérir une connaissance de la situation en temps opportun sans sacrifier l'exactitude des renseignements.

Perspectives : Nous avons pour objectif d'intégrer tôt ou tard sous forme de service les capacités de GOSSIP aux outils d'analyse actuels.