
Analyse de grands graphes de terrain

Exposé invité - CORIA 2008

Matthieu Latapy

*LIP6 (CNRS et UPMC)
174 avenue du président Kennedy
75016 Paris
Matthieu.Latapy@lip6.fr*

Dans de nombreux contextes, il est naturel et pertinent de représenter (une partie de) l'information étudiée sous forme de relations entre objets ; parfois l'information disponible se réduit même essentiellement à un tel ensemble de relations. Dans tous les cas, on se retrouve confronté à des graphes qui ne sont pas issus d'une modélisation mathématisée, mais qui représentent des données réelles, issues d'un terrain particulier ; c'est pourquoi on les appelle graphes de terrain. Les exemples sont innombrables, mais en informatique on parle souvent de la topologie de l'internet (infrastructures physiques ou logiques), de graphes issus du web (blogs par exemple), de graphes d'échanges (pair-à-pair par exemple)... Dans d'autres disciplines, on parle de réseaux sociaux, biologiques, linguistiques...

Ces grands graphes capturent généralement une information extrêmement riche sur les objets associés et le terrain sous-jacent. Une des problématiques essentielles du domaine est donc d'être capable d'identifier cette information, de distinguer l'information pertinente, et de l'extraire des graphes disponibles. Plus globalement, la simple description de ces graphes (à quoi ressemblent-ils ?) est un défi en soi ; c'est ce qu'on appelle l'analyse des graphes de terrain.

Depuis la fin des années 90, une intense activité de recherche s'est développée sur ces objets (suite à deux facteurs complémentaires : la découverte du fait qu'ils ont des propriétés non-triviales en commun, permettant de les envisager comme un tout ; et la disponibilité de plus en plus généralisée de données massives à étudier). Aujourd'hui, un ensemble de propriétés relativement large est disponible et communément accepté comme point de départ pertinent. Cet ensemble reste toutefois largement insuffisant au regard des besoins et des attentes, et il reste à étendre. De plus, l'acquis concerne essentiellement des graphes simples, c'est à dire constitués uniquement de

nœuds et de liens. Si on ajoute une orientation aux liens, un poids sur les liens, différents types de nœuds, une dynamique temporelle... alors l'existant reste extrêmement pauvre. Pourtant, en pratique, les graphes mis en jeu ont le plus souvent plusieurs de ces caractéristiques.

Il est bien clair que les méthodes d'apprentissage, de fouille de données, de classification, et plus généralement de recherche d'information, ont beaucoup à apporter à ce domaine d'étude. Bien que des contributions significatives aillent déjà dans ce sens, il reste beaucoup à faire pour que les deux mondes se comprennent mutuellement, tant en ce qui concerne les problématiques soulevées, les méthodes et outils disponibles, et l'interprétation des résultats. Cet exposé s'inscrit dans cette perspective, en proposant une présentation succincte du domaine vu sous l'angle des graphes, et en détaillant certaines problématiques sur lesquelles un apport de domaine de la recherche d'information semble particulièrement pertinent.