I *Knowledge Graph* sono un possibile approccio per costruire grandi basi di conoscenza, ovvero collezioni strutturate di dati interpretabili. Un esempio di grafi di conoscenza (*Knowledge Graph*) distribuiti sono i *Linked Data*, che potrebbero avere un impatto importante nella nascita di una "*Linked Data Economy*".

di Corrado Randaccio (randaccio@pubblimatica.it)

I KNOWLEDGE GRAPH PER COSTRUIRE GRANDI COLLEZIONI STRUTTURATE DI DATI

***Corrado RANDACCIO** svolge attività didattica universitaria nei CdL e Scuole di Specializzazione come professore a contratto di Informatica presso l'Università degli Studi di Milano e Cagliari, Facoltà di Medicina e Chirurgia, Scuola di Specializzazione in Ginecologia ed Ostetricia e l'Università degli Studi di Cagliari, Facoltà di Scienze della Formazione. Corso di Laurea in Scienze e Tecniche Psicologiche. E' docente per le discipline di: Sistemi Informativi e Basi Dati della Pubblica Amministrazione, E-Government, Telelavoro, RUPA (Rete Unitaria Pubblica Amministrazione), Diritto di accesso alla documentazione e Privacy nei corsi di riqualificazione del personale della Pubblca Amministrazione Locale e ASL.*

1. *Introduzione*

Il problema di rappresentare la conoscenza in modo che un computer possa elaborarla direttamente è ben noto nel campo dell'Intelligenza Artificiale ed è stato studiato a partire dal 1970 circa. Un approccio basato su grafi (strutture matematiche rappresentate da un insieme di nodi, o vertici, che possono essere collegati tra loro da archi) per la rappresentazione della conoscenza è stato discusso da diverso tempo, a partire dall'introduzione dei grafi concettuali[1].

I *Knowledge Graph* sono in questo senso, un possibile approccio per costruire grandi basi di conoscenza, ovvero collezioni strutturate di fatti relativi a uno stato del mondo, una rete di dati collegati appartenenti a un dominio (che costituisce il contesto di partenza), collegato a sua volta ad altri set di dati esterni, ovvero fuori dal dominio, in un contesto di relazioni sempre più estese. Le domande che ci poniamo riguardo ai sistemi complessi organizzati in rete ruotano tutte intorno a un unico quesito chiave: esistono regole rappresentabili matematicamente che ci permettono di capire come si potrà comportare una rete (per sua natura complessa, imprevedibile, in continua evoluzione dinamica) andando a studiare e a categorizzare la sua struttura fisica, ovvero il modo in cui sono connessi i nodi?

Dalla risposta a questa domanda derivano poi a cascata le domande relative alla definizione dei parametri misurabili che descrivono la rete e le proprietà che emergono in relazione alle diverse configurazioni del sistema reticolare, sia esso biologico, ecologico, sociale o digitale. E derivano le domande "tecniche" relative alle possibili modalità di gestire i fenomeni complessi che nascono e si sviluppano nella rete stessa. Si possono comprendere alcune specifiche tipologie di comportamento di un sistema complesso affrontandolo dal lato topologico: dal punto di vista strutturale, infatti, esso è formato da grandi quantità di elementi e di sotto-sistemi connessi e interagenti gli uni con gli altri in maniera non lineare; e quindi può essere modellizzato come una rete o un grafo, utilizzando le leggi della geometria e della matematica discreta che sono state messe a punto negli ultimi due decenni. Negli anni si sono individuate varie grandezze di misura della rete (dalla lunghezza caratteristica o *path lenght*, al *clustering coefficient*) che permettono di definire il tipo di struttura del sistema e quindi di prevedere come questo reagirà a sollecitazioni interne o esterne.

Si sono così individuati i profili delle reti casuali (*random*) e di quelle ordinate (le più semplici), delle reti di particolare efficienza nella trasmissione delle informazioni dette small world, e delle reti organizzate a grappoli (*cluster*) più o meno imponenti, fino alle forme più caratteristiche per le reti sociali, ovvero quelle cosiddette scale free, dotate di pochi enormi hub e di moltissimi nodi con pochi link. A seconda della tipologia di rete che struttura il sistema e a seconda dei valori che assumono le sue grandezze caratteristiche, è possibile misurare diverse proprietà preziose del sistema, come la sua fragilità o la sua resistenza di fronte a un attacco esterno o a un crash interno, o come la velocità di diffusione di un segnale o di un contagio epidemico. Conoscere e controllare le rappresentazioni simboliche di base (semplici algoritmi di matematica discreta) di questi modelli significa "acquisire una protesi cognitiva" che ci fornisce strumenti efficaci per affrontare fenomeni e problemi complessi in ambienti turbolenti, caotici e ricchi di miniere nascoste come quello dei *Big Data*.

2. *La visione memetica*

Le cose che abbiamo detto a proposito della topologia delle reti e delle loro proprietà emergenti, misurabili grazie a parametri come il *clustering coefficient* o il *power low exponent*, sono applicabili a un contesto ancora poco definito ma di estremo interesse: la "memetica" e la sua estensione nel Web, la "tecnomemetica".

Il primo termine, coniato da Richard Dawkins nel 1976, in assonanza con il termine "genetica", deriva dall'osservazione sperimentale che idee, comportamenti, mode, credenze e religioni (ovvero i memi) si estinguono o si affermano nell'ambiente sociale secondo il setaccio della selezione naturale, come fanno i geni negli organismi viventi. Ciò significa che possiamo applicare ai memi le stesse leggi di diffusione che abbiamo individuato nei modelli delle reti sociali o informatiche. Inoltre possiamo studiare

1 John F. Sowa. 1976. Conceptual graphs for a data base interface. IBM Journal of Research and Development 20, 4 (July 1976), 336-357.

la struttura topologica della rete per capire le dinamiche di base del propagarsi di mode o di idee, in questo nostro mondo che ha accelerato esponenzialmente i processi memetici, prima attraverso radio e televisione e ora anche attraverso i *social network* e la telefonia *smart*. Proprio in questo ambiente tecnologico più recente sono nati e stanno proliferando i tecnomemi, replicatori di informazioni in grado di agire e replicarsi in rete senza l'aiuto dell'uomo; il termine (in inglese spesso abbreviato in "*teme*") è stato proposto da Susan Blackmore in una conferenza TED[2] del 2008 e da allora si è diffuso in molti studi di sociologia, tra accuse di nebulosità e affermazioni di utilità oggettiva.
Sempre più utilizzata nel concreto è intanto la memetica, che vediamo applicata come protesi cognitiva efficace per capire i flussi di informazione e il sentiment in rete: così per esempio la memetica fornisce già da qualche anno[3] gli strumenti per pilotare campagne di marketing virale o per prevedere i trend di un target specifico, come abbiamo visto nelle previsioni sugli andamenti della Borsa in base al sentiment che emerge da *Twitter*; e tutto questo grazie all'utilizzo di modelli e algoritmi messi a punto per lo studio di reti complesse di tipo ecologico, epidemiologico ed economico.

3. *I Linked Data*

Un esempio di grafi di conoscenza (*Knowledge Graph*) distribuiti sono i *Linked Data*. I *Linked Data* consentono di pubblicare dati e conoscenza sul Web rappresentando in modo esplicito le relazioni, e permettendo ad un computer di accedervi in modo diretto e di interrogare in modo semantico questi grafi di conoscenza distribuiti. Il concetto di *Linked Data* è strettamente connesso al web semantico, seppure il web semantico non si risolva nel solo tecnicismo dei *Linked Data*, ma richieda, per la sua costruzione, il rispetto di alcune importanti regole finalizzate alla creazione di uno strato di contenuti accessibili a processi automatizzati. I Linked Data rendono espliciti i significati e le connessioni implicitamente contenuti (o in alcuni casi, assenti) nelle risorse del web (dati, pagine, programmi etc.). *I Linked Data* (dati connessi) rappresentano un paradigma di accesso ai dati che si è diffuso negli ultimi anni e che ha lo scopo di facilitare l'accesso all'informazione da parte di agenti automatizzati (software). I *Linked Data* (spesso abbreviato in LD) hanno un potenziale di diffusione del tutto analogo a quello del *World Wide Web* in quanto sfruttano il medesimo principio, l'interconnessione attraverso indirizzi accessibili attraverso il protocollo HTTP (*HyperText Transfer Protocol*, protocollo di trasferimento di ipertesti).
Il principio è molto semplice: così come i contenuti web possono essere connessi tra loro attraverso l'identificatore che troviamo nella barra degli indirizzi di un browser (e che troviamo visualizzato come "link" all'interno della pagina), allo stesso modo blocchi di dati possono riferire l'un l'altro associando ad essi indirizzi univoci. Prerequisito per l'accesso ad un dato in modalità LD è quindi l'associare ad esso un identificatore, un URI (*Uniform Resource Identifier*, identificatore uniforme di risorsa) e consentire il suo scaricamento attraverso il protocollo HTTP. Il termine *Linked Open Data* (LOD) aggiunge ai LD la connotazione di dati aperti, liberamente accessibili. Al fine di rendere il dato non solo accessibile, ma anche comprensibile ad un agente automatizzato (associare, cioè, una "semantica" al dato) è tuttavia necessario disporre di un modello generale per la rappresentazione dei dati, costituito dal *Resource Description Framework* (RDF). Ulteriori linguaggi, sempre espressi secondo il modello definito da RDF, consentono poi di definire schemi di dati con elevata espressività: *RDF Schema* (RDFS) e *Web Ontology Language* (OWL).
Gli schemi definiti secondo questi ultimi formalismi, tipicamente identificati col termine "ontologie", consentono di modellare strutture dati con un livello di dettaglio più fine rispetto al modello relazionale (comunemente applicato nei comuni RDBMS) e al modello XML. Inoltre, le ontologie consentono di superare l'interpretazione "a mondo chiuso" dei modelli precedenti e di realizzare strutture dati che possano essere interpretati secondo la logica "a mondo aperto" che contraddistingue il Web Semantico. I *Linked Data* e le ontologie rappresentano quindi gli strumenti per accedere e strutturare, rispettivamente, dati connessi tra loro e caratterizzati da una semantica specifica. I *Linked Data* realizzano il progetto originario di Tim Berners Lee, l'inventore del Web. Quando i dati sono collegati, diventa più facile estrarne conoscenza, anche facendo leva su informazioni pubblicate da altri. I *Linked Data* sono al cuore dei progetti più innovativi delle amministrazioni pubbliche – all'estero, la BBC e molti uffici del governo UK; in Italia, la Camera dei Deputati col supporto di *Regesta.exe*. I *Linked Data* sono anche al centro di piattaforme come *Watson* di IBM, ed elementi *linked* sono presenti in prodotti come *Siri* di Apple. Google e Facebook supportano questo approccio tramite *Schema.org* e l'*Open Graph Protocol*. Recentemente sono state promosse importanti iniziative che potrebbero avere un impatto importante nella nascita di una "*Linked Data Economy*": in primo luogo Google, che introduce sul suo portale: *Google Knowledge Graph*, una funzione di ricerca che è stata introdotta il 16 maggio 2012 mentre nella versione italiana è stata attivata il 4 dicembre 2012, e che si è rivelata dalla data della sua pubblicazione ad oggi, come una sorta di *linked closed data cloud industriale*. Google ha anche parlato dell'acquisizione di *Freebase*, uno dei nodi più importanti della *LOD cloud*; in secondo luogo, è stata stipulata una coalizione tra alcuni dei più grandi motori di ricerca (tra cui Google, Yahoo e Bing) che ha prodotto un insieme di tecnologie e incentivi economici e sociali con lo scopo di indurre i produttori di contenuti ad arricchire le proprie pagine web con *markup* semantico; infine, molte grandi organizzazioni pubbliche e private si stanno avvicinando al web dei dati, anche modificando profondamente i loro modelli di business e processi di produzione, oppure creando le proprie *closed linked data clouds* (es: molte grandi aziende farmaceutiche stanno costruendo le proprie *linked cloud* aziendali).
Un domanda che ci si pone è sicuramente quella relativa alle questioni socio-economiche e al rischio di impoverimento della cosa pubblica e al generarsi di posizioni di monopolio, che queste iniziative determinano. Il mercato europeo è caratterizzato dalla presenza di una moltitudine di PMI che rappresentano la forza trainante nell'innovazione e nella crescita economica. In questo scenario, è fondamentale riuscire a concepire una strategia oculata capace allo stesso tempo di proteggere il patrimonio culturale pubblico e di offrire incentivi alle PMI per riuscire a generare una *Linked Data Economy* vitale e sostenibile. D'altra parte, stiamo assistendo a enormi passi avanti del web dei dati, il cui valore economico viene valorizzato grazie alla massa critica (di utenti, investimenti, tecnologia, visibilità nei confronti dei media e stimolo della domanda) che le aziende leader del web sono capaci di smuovere. Sicuramente il rumore mediatico creato dal *Google Knowledge Graph* – che è ancora per la gran parte costituito da *commons* – può rappresentare una grande opportunità per le PMI, che possono sfruttare gli stessi commons per soddisfare la crescente domanda di *linked data* in settori verticali e specializzati. ©

2 http://www.ted.com/talks/susan_blackmore_on_memes_and_temes.html.

3 S. Collina e V. Simonte, "Introduzione alla memetica. La comunicazione virale", Aracne, Roma 2007.