Can Scholarly Literature and Patents be Represented in a Hierarchy of Topics Structured to Contain 20 Topics per Level? Balancing Technical Feasibility with Human Usability

Michael Edwards[1], Mahadev Dovre Wudali[2], James Callahan[3], Paul Worner[4], Jeffrey Maudal[5], Patricia Brennan[6], Julia Laurin[7] and Joshua Schnell[8]

[1]*michael.edwards@thomsonreuters.com*
[2]*mahadev.wudali@thomsonreuters.com*
[3]*jim.callahan@thomsonreuters.com*
[4]*paul.worner@thomsonreuters.com*
[5]*jeff.maudal@thomsonreuters.com*
Data Center Operations, Thomson Reuters, 610 Opperman Drive, Eagan, Minnesota 55123

[6]*patricia.brennan@thomsonreuters.com*
[7]*julia.laurin@thomsonreuters.com*
[8]*joshua.schnell@thomsonreuters.com*
Intelectual Property & Science, Thomson Reuters,
1500 Spring Garden St, Philadelphia, Pennsylvania 19130

Introduction

The Intellectual Property & Science division of Thomson Reuters curates millions of records a year covering scholarly literature (Web of Science®), patents and intellectual property (Derwent World Patent Index®) and life sciences discovery (Cortellis®). These millions of records could be connected through billions of potential relationships, such as that represented by a citing relationship between literature and patents, or by different documents that pertain to similar topics. By building these relationships using machine learning techniques we hope to unite information from different data sources to enable extraction of knowledge such that the whole is greater than the sum of the parts, with minimal human effort required.

However, connecting these documents in a meaningful way is challenging from both a technological perspective as well as a usability perspective. As shown in Figure 1, studying citation patterns among approximately 250,000 articles from the Web of Science, or 1/200 of the full data set, generates a citation graph that, while rich with information, is extremely difficult to use to understand knowledge flows.

This challenge is the focus of our presentation. For this research project, we have created a graph of the topics represented in a subset of the scholarly literature and granted patents, in order to explore ways to constrain the visualization of this topic graph to emphasize usability. While many additional research areas remain, our initial findings suggest that such constraint enables users to easily explore the knowledge graph in way that maximizes understanding while minimizing user effort.

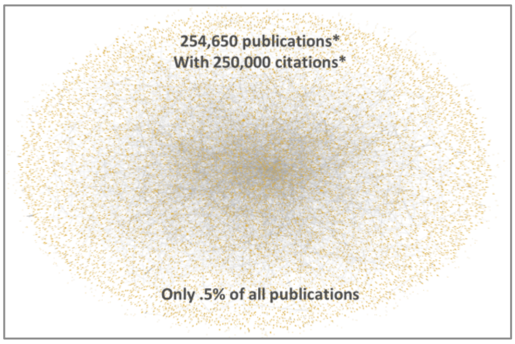


Figure 1. Ball and stick diagram of the citing relationships among a select set of publications from Web of Science®.

Generation of the Topic Graph

We chose to use topic modelling based on the latent dirichlet allocation (LDA) algorithm (Blei, Ng & Jordan, 2003) to generate connections between documents that reflect the shared knowledge among scholarly articles and granted patents. From Web of Science, we selected 27 million publications published since 1990 that had abstracts in English. Our past experience with LDA topic modelling led us to take a hierarchical approach to clustering the documents based on topics. We created a tree of over 1 million topics for the corpus, parceling out the topics into manageable chunks (20 at a glance) which were a better fit for human perception. We also created our own algorithm for applying these topics to patents, demonstrating a flexible, unsupervised technique for combining two distinct content sets. We found that the hierarchy we produced generally exhibited 4 to 5 levels of depth to the terminal nodes or documents.

Understanding the Knowledge Graph

We created the Epiphany tool to more effectively navigate the corpus of scholarly articles, using both browse and search interactions. As shown in Figure 2, the tool supports drill-down (e.g. 2.6 million articles assigned to an algorithm-focused topic; left side green), as well as search, (e.g. 8 topics strongly related to "genetic programming"; right side orange). This allows users to interact with topics and the relevant documents to understand the underlying data.

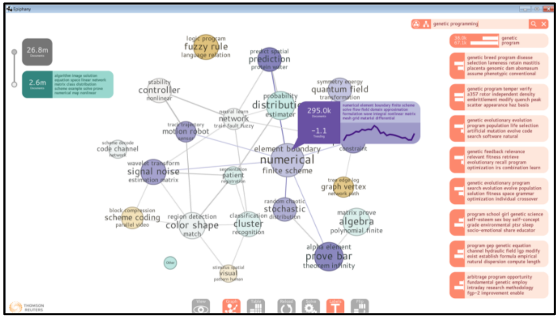
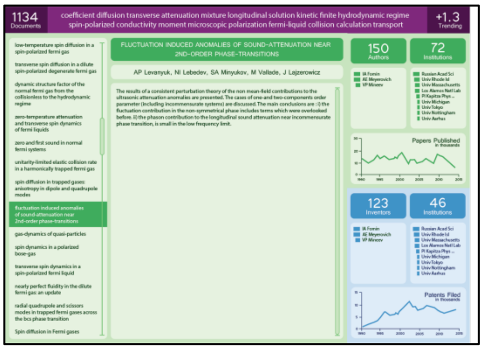

Epiphany
26.8m
Documents
2.6m
Documents
algorithm image solution equation space linear network matrix class distribution scheme example solve prove numerical map nonlinear
logic program
fuzzy rule
language relation
predict spatial
prediction
protein water
stability
controller
nonlinear
symmetry energy
quantum field
transformation
probability
distribution
estimator
neural learn
network
train fault fuzzy
track trajectory
motion robot
sensor
295.0k
Documents
-1.1
Trending
numerical element boundary finite scheme solve flow field domain approximation formulation wave integral nonlinear matrix mesh grid material differential
scheme decode
code channel
network
element boundary
numerical
finite scheme
constraint
wavelet transform
signal noise
estimation matrix
segmentation
patient
registration
tree edge log
graph vertex
network path
random chaotic
stochastic
distribution
block compression
scheme coding
parallel video
region detection
color shape
match
classification
cluster
recognition
matrix prove
algebra
polynomial finite
alpha element
prove bar
theorem infinity
Other
stimulus spatial
visual
pattern human
genetic programming
38.0k
67.1k
genetic program
genetic breed program disease selection lameness retain mastitis placenta genomic dam abomasum assume phenotypic conventional
genetic program temper verify a357 rotor independent density embrittlement modify quench peak scatter appearance haz basis
genetic evolutionary evolution program population life selection artificial mutation evolve code search software natural
genetic feedback relevance relevant fitness retrieve evolutionary recall program optimization irs combination learn
genetic evolutionary program search evolution evolve population solution fitness space grammar optimization individual crossover
program school girl genetic science self-esteem sex boy self-concept grade environmental ptsr sleep socio-emotional share educator
program gep genetic equation channel hydraulic field lgp modify exist establish formula empirical natural dispersion compute length
arbitrage program opportunity fundamental genetic employ intraday research methodology fgp-2 improvement enable
THOMSON REUTERS
View
Graph
Table
Reload
Solve
Labels
Flip

Figure 2. Screenshot of Epiphany tool showing topic clusters matching "genetic programming" search criteria.

Drilling down into the topic details is show in Figure 3. At the top in purple are statistics on the topic itself including the number of documents closely associated with the topic, the most frequent terms and the Trending metric score for the topic.



1134 Documents

coefficient diffusion transverse attenuation mixture longitudinal solution kinetic finite hydrodynamic regime spin-polarized conductivity moment microscopic polarization fermi-liquid collision calculation transport

+1.3 Trending

low-temperature spin diffusion in a spin-polarized fermi gas

transverse spin diffusion in a dilute spin-polarized degenerate fermi gas

dynamic structure factor of the normal fermi gas from the collisionless to the hydrodynamic regime

zero-temperature attenuation and transverse spin dynamics of fermi liquids

zero and first sound in normal fermi systems

unitarity-limited elastic collision rate in a harmonically trapped fermi gas

spin diffusion in trapped gases: anisotropy in dipole and quadrupole modes

fluctuation induced anomalies of sound-attenuation near 2nd-order phase-transitions

gas-dynamics of quasi-particles

spin dynamics in a polarized bose-gas

transverse spin dynamics in a spin-polarized fermi liquid

nearly perfect fluidity in the dilute fermi gas: an update

radial quadrupole and scissors modes in trapped fermi gases across the bcs phase transition

Spin diffusion in Fermi gases

FLUCTUATION INDUCED ANOMALIES OF SOUND-ATTENUATION NEAR 2ND-ORDER PHASE-TRANSITIONS

AP Levanyuk, NI Lebedev, SA Minyukov, M Vallade, J Lajzerowicz

The results of a consistent perturbation theory of the non mean-field contributions to the ultrasonic attenuation anomalies are presented. The cases of one-and two-components order parameter (including incommensurate systems) are discussed. The main conclusions are : i) the fluctuation contribution in the non-symmetrical phase includes terms which were overlooked before. ii) the phason contribution to the longitudinal sound attenuation near incommensurate phase transition, is small in the low frequency limit.

150 Authors

72 Institutions

IA Fomin
AE Meyerovich
VP Mineev

Russian Acad Sci
Univ Rhode Isl
Univ Massachusetts
Los Alamos Natl Lab
PI Kapitza Phys ...
Univ Michigan
Univ Tokyo
Univ Nottingham
Univ Aarhus

123 Inventors

46 Institutions

IA Fomin
AE Meyerovich
VP Mineev

Russian Acad Sci
Univ Rhode Isl
Univ Massachusetts
Los Alamos Natl Lab
PI Kapitza Phys ...
Univ Michigan
Univ Tokyo
Univ Nottingham
Univ Aarhus

Figure 3. Screenshot of Epiphany tool Topic Details screen.

The right side of the panel contains two statistics sections, one in green for scientific papers and one in blue for patents. The header for each of the sections includes counts of the unique number of authors (or inventors) and unique number of institutions (or assignees) responsible for creation of the documents associated with the topic. Below these counts are a breakdown of the most commonly mentioned authors (inventors) and institutions (assignees). Finally, the bottom part of the statistics section is a graph of the proportion of documents assigned to this topic out of all documents published for each year.

Project Outcomes

The purpose of this research project is to test the application of scalable machine learning techniques to generate a knowledge graph that is accessible to the analyst. Now that we have developed the Epiphany tool, we have begun using it to gather feedback on this approach from a cross section of potential users. We expect to present that feedback at the ISSI2015 conference specifically to answer the question of whether a topic graph of millions of records of scholarly literature and granted patents can indeed be represented in hierarchical structure with a maximum of 20 topics at each level.

Acknowledgments

We acknowledge the support of the Intellectual Property & Science staff and the Data Center Operations staff for improvements made to this research project.

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.