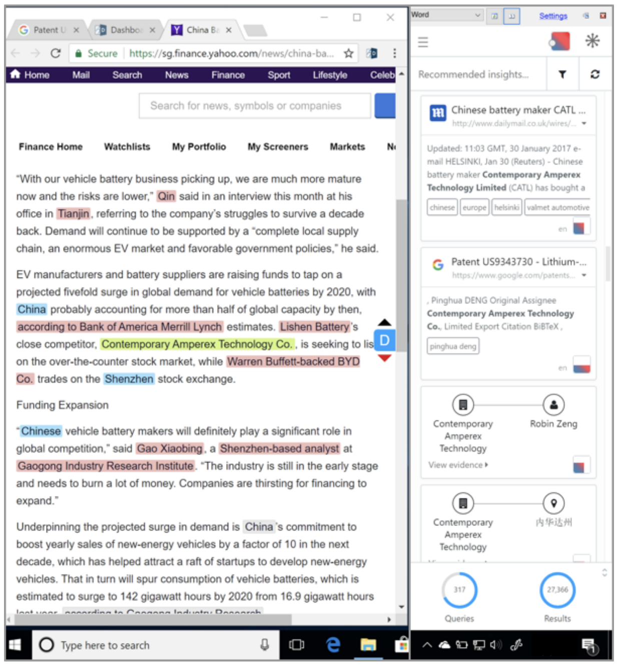
How it works: collaborative intelligence.

1. Diffeo

The Diffeo platform uses collaborative search and discovery algorithms to augment the intelligence of human users.

Users invite Diffeo into their working documents, so that Diffeo can highlight knowledge gaps and uncover connections[1]. Instead of manually searching, users collaborate with machines to quickly see the whole story, so they can focus on analysis and action.

Diffeo uses natural language processing to recognize mentions of people, organizations and other concepts in PDFs and pages in your web browser and draft documents in Word, Outlook, Quip etc. By inviting Diffeo into these familiar user interfaces, they become ***user-facing sensors*** that enable the machine to understand your current work. Diffeo builds a dynamic knowledge graph of related entities by mining many data sources, including your Exchange email and team shared drives, and automatically formulates queries to lower-level search engines.



"With our vehicle battery business picking up, we are much more mature now and the risks are lower," Qin said in an interview this month at his office in Tianjin, referring to the company's struggles to survive a decade back. Demand will continue to be supported by a "complete local supply chain, an enormous EV market and favorable government policies," he said.

EV manufacturers and battery suppliers are raising funds to tap on a projected fivefold surge in global demand for vehicle batteries by 2020, with China probably accounting for more than half of global capacity by then, according to Bank of America Merrill Lynch estimates. Lishen Battery's close competitor, Contemporary Amperex Technology Co., is seeking to lis on the over-the-counter stock market, while Warren Buffett-backed BYD Co. trades on the Shenzhen stock exchange.

Funding Expansion

"Chinese vehicle battery makers will definitely play a significant role in global competition," said Gao Xiaobing, a Shenzhen-based analyst at Gaogong Industry Research Institute. "The industry is still in the early stage and needs to burn a lot of money. Companies are thirsting for financing to expand."

Underpinning the projected surge in demand is China's commitment to boost yearly sales of new-energy vehicles by a factor of 10 in the next decade, which has helped attract a raft of startups to develop new-energy vehicles. That in turn will spur consumption of vehicle batteries, which is estimated to surge to 142 gigawatt hours by 2020 from 16.9 gigawatt hours

Chinese battery maker CATL ...
http://www.dailymail.co.uk/wires/...

Updated: 11:03 GMT, 30 January 2017 e-mail HELSINKI, Jan 30 (Reuters) - Chinese battery maker **Contemporary Amperex Technology Limited** (CATL) has bought a

chinese | europe | helsinki | valmet automotive

en

Patent US9343730 - Lithium-...
https://www.google.com/patents...

, Pinghua DENG Original Assignee **Contemporary Amperex Technology Co.**, Limited Export Citation BiBTeX ,

pinghua deng

en

Contemporary Amperex Technology — Robin Zeng

View evidence ▸

Contemporary Amperex Technology — 内华达州

317 Queries

27,366 Results

Figure 1: Diffeo recommends related concepts by inserting red highlights on pages in your browser (left), and by presenting cards in the list (right). By discarding or acting on these suggestions, the user's natural workflow drives Diffeo's learning loops that iteratively expand the knowledge graph for this project.

Diffeo Labs has been studying collaboration as a cornerstone of intelligence since 2012[2]. The recommender engine grew out of DARPA Memex and NIST's Text Retrieval Conference (TREC) where the Diffeo team organized the TREC Knowledge Base Acceleration (KBA) and Dynamic Domain (DD) evaluations from 2012 through 2015.

To ***measure*** which recommender algorithms suggest novel content to users, Diffeo has refined protocols for both offline evaluations and task completion studies. In 2016, the Diffeo recommender's results on these metrics caused it to graduate from Diffeo Labs into production use. Today, Diffeo Labs is studying how to build machines that collaborate with people[3].

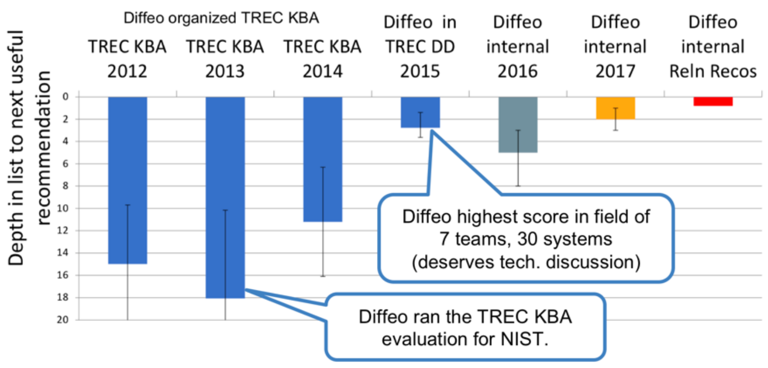


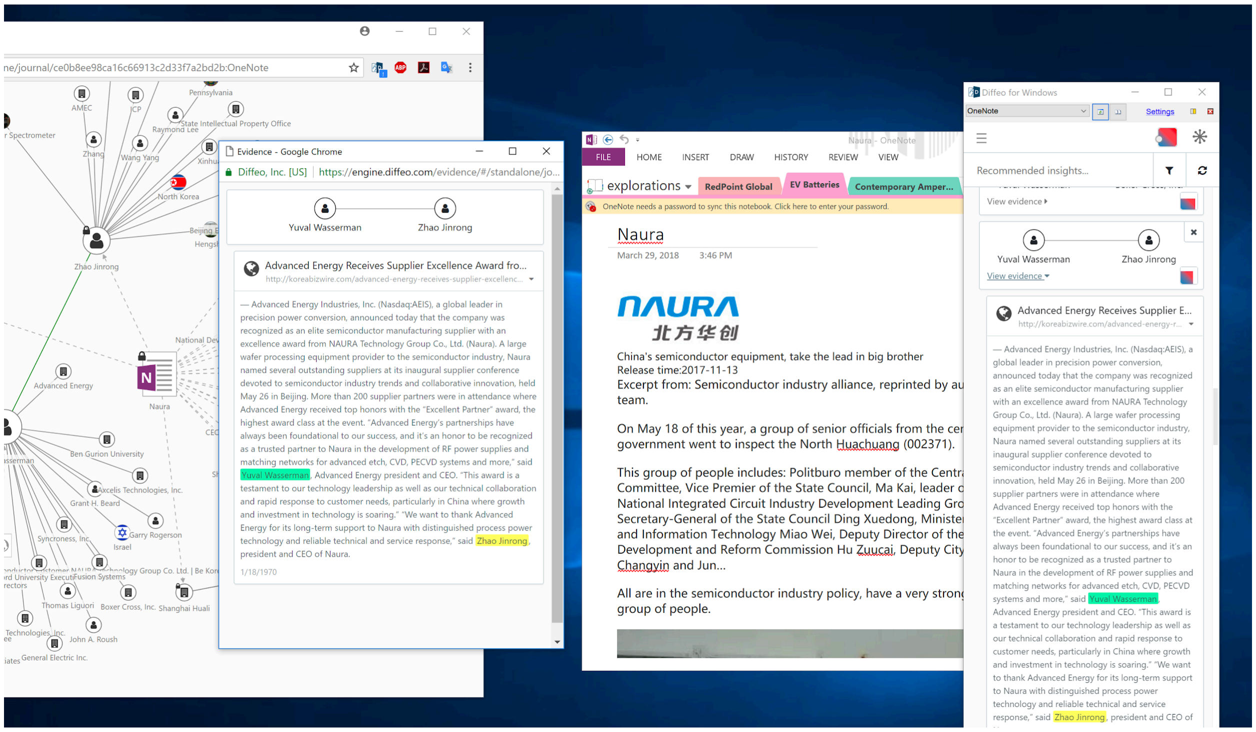
Figure 2: In 2012, the best systems presented fourteen unhelpful items on average before uncovering a citation-worthy document. In 2013, we doubled task size to 1.2 billion documents, the largest-ever stream corpus released for public evaluations. Systems performed worse on this harder task. In 2014, systems started removing world knowledge, which improved recommendations. In 2015, Diffeo's ***user-facing sensors*** enabled dynamic reranking to show surprising and *novel* connections.

[1] US Patent 9.275,132. Entity-Centric Knowledge Discovery. https://libpatent.com/patents/09275132

[2] See papers at http://trec-dd.org/2015/2015.html and http://trec-kba.org

[3] **Kleiman-Weiner, M. (Diffeo Chief Scientist)**, Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. *Conference of the Cognitive Science Society*. [pdf]

1.1. Related Entities with Evidence



— Advanced Energy Industries, Inc. (Nasdaq:AEIS), a global leader in precision power conversion, announced today that the company was recognized as an elite semiconductor manufacturing supplier with an excellence award from NAURA Technology Group Co., Ltd. (Naura). A large wafer processing equipment provider to the semiconductor industry, Naura named several outstanding suppliers at its inaugural supplier conference devoted to semiconductor industry trends and collaborative innovation, held May 26 in Beijing. More than 200 supplier partners were in attendance where Advanced Energy received top honors with the "Excellent Partner" award, the highest award class at the event. "Advanced Energy's partnerships have always been foundational to our success, and it's an honor to be recognized as a trusted partner to Naura in the development of RF power supplies and matching networks for advanced etch, CVD, PECVD systems and more," said Yuval Wasserman, Advanced Energy president and CEO. "This award is a testament to our technology leadership as well as our technical collaboration and rapid response to customer needs, particularly in China where growth and investment in technology is soaring." "We want to thank Advanced Energy for its long-term support to Naura with distinguished process power technology and reliable technical and service response," said Zhao Jinrong, president and CEO of Naura.

1/18/1970

Naura

March 29, 2018 3:46 PM

China's semiconductor equipment, take the lead in big brother
Release time:2017-11-13
Excerpt from: Semiconductor industry alliance, reprinted by au
team.

On May 18 of this year, a group of senior officials from the cen
government went to inspect the North Huachuang (002371).

This group of people includes: Politburo member of the Centra
Committee, Vice Premier of the State Council, Ma Kai, leader o
National Integrated Circuit Industry Development Leading Gro
Secretary-General of the State Council Ding Xuedong, Ministe
and Information Technology Miao Wei, Deputy Director of the
Development and Reform Commission Hu Zuucai, Deputy City
Changyin and Jun...

All are in the semiconductor industry policy, have a very stron
group of people.

— Advanced Energy Industries, Inc. (Nasdaq:AEIS), a global leader in precision power conversion, announced today that the company was recognized as an elite semiconductor manufacturing supplier with an excellence award from NAURA Technology Group Co., Ltd. (Naura). A large wafer processing equipment provider to the semiconductor industry, Naura named several outstanding suppliers at its inaugural supplier conference devoted to semiconductor industry trends and collaborative innovation, held May 26 in Beijing. More than 200 supplier partners were in attendance where Advanced Energy received top honors with the "Excellent Partner" award, the highest award class at the event. "Advanced Energy's partnerships have always been foundational to our success, and it's an honor to be recognized as a trusted partner to Naura in the development of RF power supplies and matching networks for advanced etch, CVD, PECVD systems and more," said Yuval Wasserman, Advanced Energy president and CEO. "This award is a testament to our technology leadership as well as our technical collaboration and rapid response to customer needs, particularly in China where growth and investment in technology is soaring." "We want to thank Advanced Energy for its long-term support to Naura with distinguished process power technology and reliable technical and service response," said Zhao Jinrong, president and CEO of

Figure 3: related entity recommendations with evidence bundles.

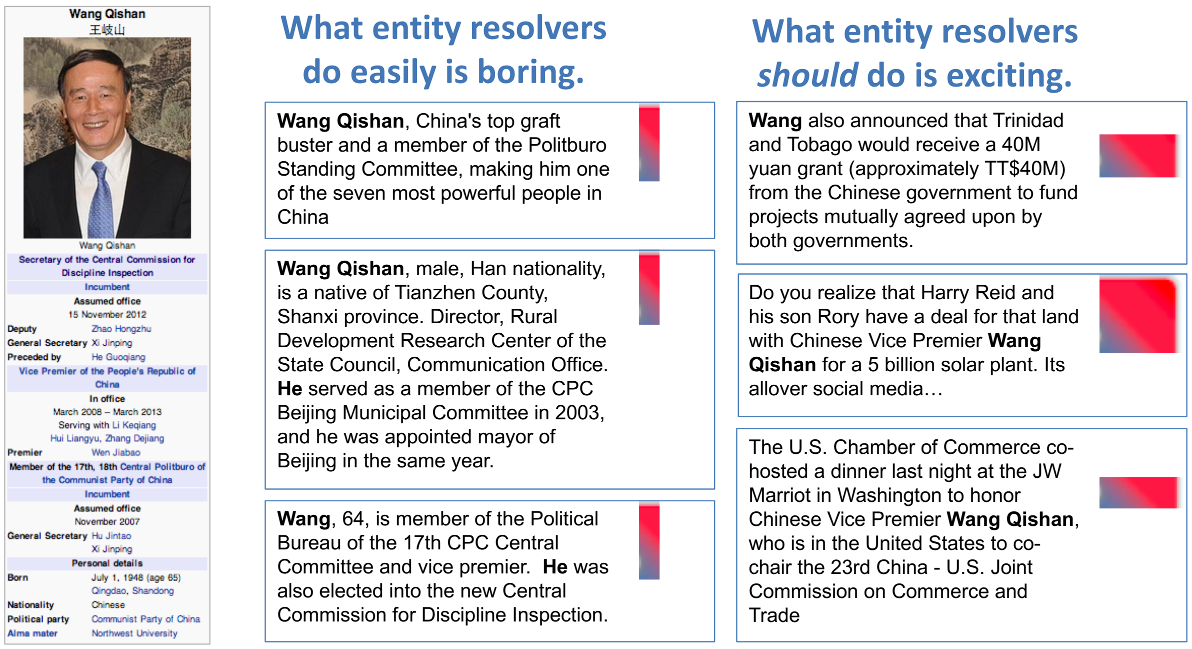
A central capability of the system is relation recommendation. Based on the people and companies that Diffeo sees you studying, it attempts to suggest related people and companies that are relevant to your analysis and *that you have not yet considered.* By presenting these related entities with evidence of the connection to your current work, the system attempts to expand your perspective. Figure 3 shows two different visual presentations of a relationship between two people, Yuval Wasserman and Zhao Jinrong. On the right, the relationship is presented alongside a document that the user is writing about Naura Semiconductor and its CEO Zhao Jinrong in OneNote. On the left, the relationship is presented as an edge between two concept vertices in a Diffeo Knowledge Board, which is a collaborative interactive workspace. In both cases, a document substantiating the relationship is presented.

in Figure 2, the red-colored evaluation result describes evidence bundles. This measurement counts each document in evidence bundle as a true positive if the user accepts a related entity recommendation. This allows a single user action to count for more than one citation.

1.2. Coref Decision Boundary

In 2011, before Diffeo was founded, John Frank was the Chief Architect for Search at Nokia. He teamed up with two other Hertz Fellows, Max Kleiman-Weiner and Dan Roberts, to help NIST develop the TREC KBA track to evaluate algorithms for helping Wikipedians write Wikipedia faster.

The prevailing wisdom at the time was that if someone could build a sufficiently large-scale entity disambiguation engine, then analysts would get much more value from HLT. Disambiguation is a technical name for the fundamental step in human communication in which an observer recognizes the meaning of a phrase or utterance written or spoken by another person. Operationally, recognizing a mention of a concept or entity means identifying other mentions that also refer to that concept. One says that the two mentions are co-referent, i.e. refer to the same thing. Algorithms for recognizing meaning take the same operational approach, and are often called coreference resolution algorithms, or "coref" for short.



Wang Qishan
王岐山

Wang Qishan

Secretary of the Central Commission for Discipline Inspection	
Incumbent	
Assumed office 15 November 2012	
Deputy	Zhao Hongzhu
General Secretary	Xi Jinping
Preceded by	He Guoqiang
Vice Premier of the People's Republic of China	
In office March 2008 – March 2013 Serving with Li Keqiang Hui Liangyu, Zhang Dejiang	
Premier	Wen Jiabao
Member of the 17th, 18th Central Politburo of the Communist Party of China	
Incumbent	
Assumed office November 2007	
General Secretary	Hu Jintao Xi Jinping
Personal details	
Born	July 1, 1948 (age 65) Qingdao, Shandong
Nationality	Chinese
Political party	Communist Party of China
Alma mater	Northwest University

What entity resolvers do easily is boring.

Wang Qishan, China's top graft buster and a member of the Politburo Standing Committee, making him one of the seven most powerful people in China

Wang Qishan, male, Han nationality, is a native of Tianzhen County, Shanxi province. Director, Rural Development Research Center of the State Council, Communication Office. **He** served as a member of the CPC Beijing Municipal Committee in 2003, and he was appointed mayor of Beijing in the same year.

Wang, 64, is member of the Political Bureau of the 17th CPC Central Committee and vice premier. **He** was also elected into the new Central Commission for Discipline Inspection.

What entity resolvers *should* do is exciting.

Wang also announced that Trinidad and Tobago would receive a 40M yuan grant (approximately TT$40M) from the Chinese government to fund projects mutually agreed upon by both governments.

Do you realize that Harry Reid and his son Rory have a deal for that land with Chinese Vice Premier **Wang Qishan** for a 5 billion solar plant. Its allover social media…

The U.S. Chamber of Commerce co-hosted a dinner last night at the JW Marriot in Washington to honor Chinese Vice Premier **Wang Qishan**, who is in the United States to co-chair the 23rd China - U.S. Joint Commission on Commerce and Trade

An entity profile used as a retrieval target for generating recommendations.

High confidence, low novelty recommendations typical of a coref classifier.

High novelty recommendations, often near the decision boundary of the coref classifier.

Figure 4: results from Diffeo's hierarchical clustering system, called Bigforest, which groups together mentions of entities based on the likelihood that they refer to the same entity.

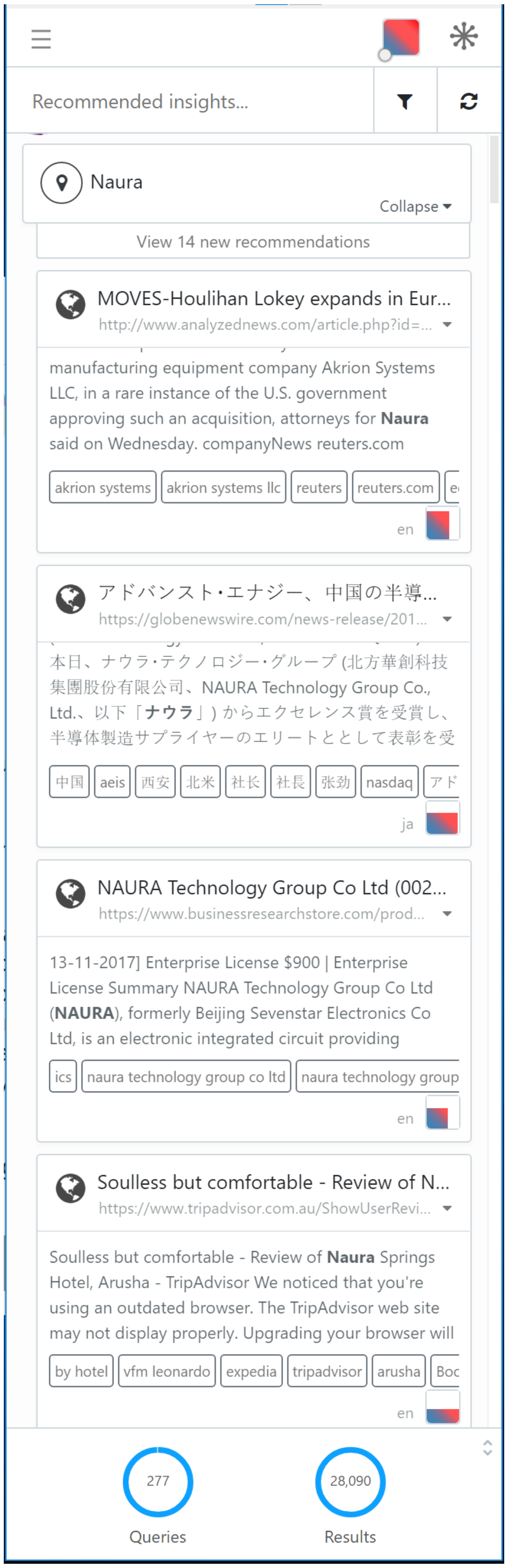
The red-to-blue boxes in Figure 4 are called "novelty-confidence indicators." The *height* of each indicator is the coreference confidence, i.e. the system's estimate that this text refers to the same entity that the user is studying. The *width* indicates the likelihood that the user will discover new information when they open this document, i.e. the number of useful red-colored highlights that the user might wish to use (see *starring* in section 3.1).

Helping analysts reach the decision boundary accelerates discovery.

The left-hand set of results in Figure 4 is ranked by confidence. Each result shows a tall-and-skinny blue-to-red box indicating that the system is highly confident that this snippet refers to Wang Qishan, the Chinese leader whose Wikipedia article is shown on the left.

When such a confidence-ordered set of results is shown to a research analyst, they typically discarded these results—even though they are correct! When asked why, analysts say, "I already knew that. It wasn't novel."

The right-hand set of results in Figure 4 have wider red-to-blue boxes. These documents contain more mentions of new relations, new topics, and potential surprises. Some boxes are shorter, meaning that the system is less confident that the mention is coreferent to the user's entity of interest. For example, consider the upper-right example. Is this the same person visiting Caribbean nations to give out money?



Naura

MOVES-Houlihan Lokey expands in Eur...
http://www.analyzednews.com/article.php?id=...
manufacturing equipment company Akrion Systems LLC, in a rare instance of the U.S. government approving such an acquisition, attorneys for **Naura** said on Wednesday. companyNews reuters.com
akrion systems | akrion systems llc | reuters | reuters.com
en

アドバンスト・エナジー、中国の半導...
https://globenewswire.com/news-release/201...
本日、ナウラ·テクノロジー·グループ (北方華創科技集團股份有限公司、NAURA Technology Group Co., Ltd.、以下「**ナウラ**」) からエクセレンス賞を受賞し、半導体製造サプライヤーのエリートとして表彰を受
中国 | aeis | 西安 | 北米 | 社长 | 社長 | 张劲 | nasdaq | アド
ja

NAURA Technology Group Co Ltd (002...
https://www.businessresearchstore.com/prod...
13-11-2017] Enterprise License $900 | Enterprise License Summary NAURA Technology Group Co Ltd (**NAURA**), formerly Beijing Sevenstar Electronics Co Ltd, is an electronic integrated circuit providing
ics | naura technology group co ltd | naura technology group
en

Soulless but comfortable - Review of N...
https://www.tripadvisor.com.au/ShowUserRevi...
Soulless but comfortable - Review of **Naura** Springs Hotel, Arusha - TripAdvisor We noticed that you're using an outdated browser. The TripAdvisor web site may not display properly. Upgrading your browser will
by hotel | vfm leonardo | expedia | tripadvisor | arusha | Boc
en

As the Diffeo team continued to shadow analysts, we observed that they would discard recommendations until they got deep enough in the list that they found surprising results. This naturally comes with some false positives, so the user experience must allow users to efficiently skip over noise — and further, the machine must learn from those actions.

In machine learning, this region of the data is often called the **decision boundary**. For a human learning about something new, the decision boundary moves. To keep up with users, Diffeo has engineered the recommender engine to track with the user, to move with the user's decision boundary.

1.3. Automatic Gap Analysis

Information retrieval systems generally presented ranked lists rather than sets. Metrics for evaluating ranked lists are fundamentally different from set-based measures (see Section 3.3.5), and are the subject of extensive research in TREC and other information retrieval forums, like SIGIR.

Through the experience building ground truth data for TREC KBA, Diffeo found that user's retrieval objectives are distinctly different from the objective functions that standard disambiguation algorithms optimize. Entity disambiguation algorithms seek to optimize the probability that two mentions are coreferent, which is correlated with overlapping contexts. A user studying an entity seeks to optimize the likelihood that the next text s/he reads carries new information about an entity of interest, which means *finding contexts that do not overlap with contexts the user already has.*

These two competing factors motivated the two-dimensional red-to-blue colored indicators on recommendation cards— see examples in Figure 5. At the top of the queue of recommendations is a two-dimensional filter control with a slider that allows the user to move the thresholds for showing recommendation cards.

Metrics for measuring the effectiveness of ranking function are the subject of TREC Dynamic Domain (DD) and simulation harness that we built for it. The core idea in TREC DD is the notion of a subtopic as a discrete aspect of a subject area that a user is trying to understand. By recording relevance judgments that assign documents and passages to subtopics within a query topic, the TREC DD protocol enables us to evaluate a system's ability to uncover those aspects of a topic that the user does not yet know.

The Diffeo platform includes several algorithms for uncovering missing subtopics including relation recommendations (Section 1.1) and cross-language entity discovery (Section 3.3.4).

2. Collaborative Data Integration

Diffeo lets users stay in their work without needing to jump into dozens of disparate tools to find content. The recommendations tap into vast corporate archives and third-party data stores so users avoid "getting lost in the Internet," guessing at search terms to find novel or critical info.

As shown in Figure 6, Diffeo Engine consists of a stack of docker containers that can be deployed to a wide range of private cloud environments, including Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. Diffeo's SaaS Platform (DSP) is operated in GCP at engine.diffeo.com.

Diffeo's worker containers and storage system can be scaled horizontally, and routinely handle hundreds of millions of documents and thousands of concurrent users.

2.1. User's Files

Users can invite Diffeo into private data sources, such as Microsoft Exchange, Windows Shared Drives, Google Drive[1], Dropbox[1], etc. Figure 4 shows how Diffeo joins an Active Directory domain to enable users to "Opt-In" to Exchange email indexing via kerberos delegation.

[1] Today, documents in cloud drives are searchable with Diffeo's smart tag search experience, and a product upgrade planned for H2 2018 will make them accessible to the recommender engine.

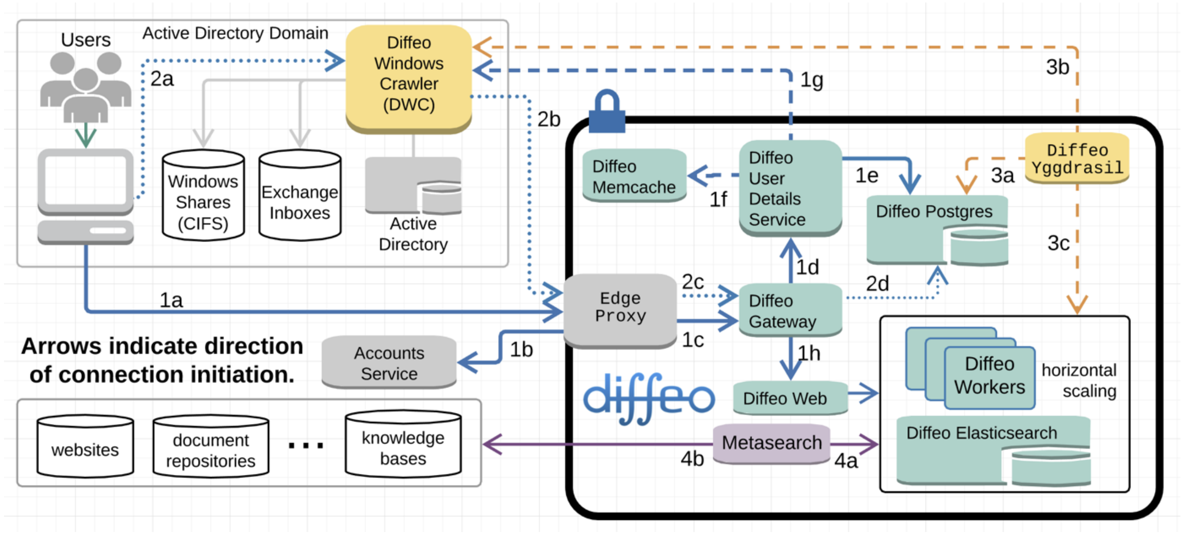


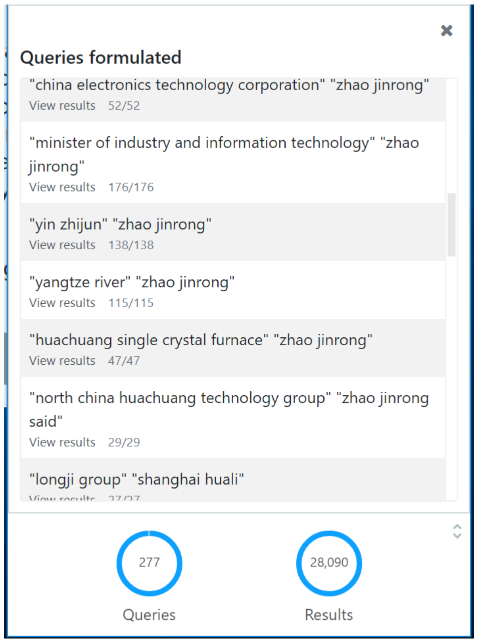
Figure 6: Diffeo Windows Crawler

2.2. Automatic Query Formulation (AQF)

As users work, Diffeo builds queries and sends them through Diffeo's metasearch infrastructure to lower-level search engines. This enables just-in-time indexing (JITI). Customers running Diffeo Engines in their private clouds use AQF to integrate Diffeo with massive data lakes and internal data portals.

As shown in the table below, `diffeo.com` has a growing number of metasearch connections described in the table below. Further details on current data integrations are available under NDA.

Name	**URL**	**Type**
Small Business Innovation Research	sbir.gov	Deep Web
Securities and Exchange Commission	sec.gov	Deep Web
MEMRI	memri.org	Deep Web
Nuclear Threat Initiative	nti.org	Deep Web
SAM	sam.gov	Deep Web
6 more significant deep Web sources for research		
Moreover LexisNexis	moreover.com	Subscription
Reddit	reddit.com	Social Media
ResearchGate	researchgate.net	Social Media
Twitter	twitter.com	Social Media
16 more major social media platforms internationally		
Bing	bing.com	Surface Web
Google	google.com	Surface Web
5 more surface Web search engines internationally		


Queries formulated

"china electronics technology corporation" "zhao jinrong"
View results 52/52

"minister of industry and information technology" "zhao jinrong"
View results 176/176

"yin zhijun" "zhao jinrong"
View results 138/138

"yangtze river" "zhao jinrong"
View results 115/115

"huachuang single crystal furnace" "zhao jinrong"
View results 47/47

"north china huachuang technology group" "zhao jinrong said"
View results 29/29

"longji group" "shanghai huali"

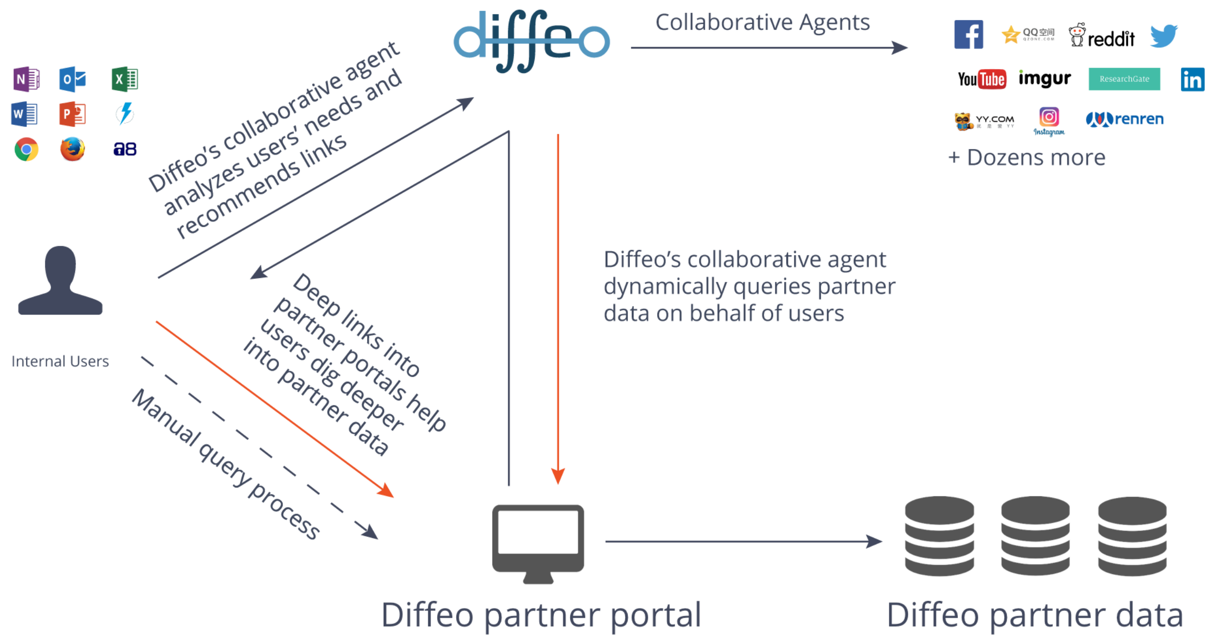
277
Queries

28,090
Results

2.3. Integrating Subscription Content

Diffeo's recommender technology offers a unique approach to integrating subscription content. By granting the Diffeo agent privileged access to their content holdings, a publisher can allow more users to ***discover*** their content. Without Diffeo, publishers generally have only two choices: (1) charge customers for full access to a content collection, or (2) charge "by the drink" for each piece of data. In both cases, the publisher is dependent on the end user to dig into the publisher's custom search portal just to find content.

Diffeo improves both modes of access. For users with full subscription access, the Diffeo agent helps ensure that they find the critical content for which they are paying. For users without a subscription, the Diffeo agent helps users efficiently discover that the publisher has content they need



diffeo
Collaborative Agents
reddit
imgur
ResearchGate
YY.COM
renren
+ Dozens more
Diffeo's collaborative agent analyzes users' needs and recommends links
Deep links into partner portals help users dig deeper into partner data
Diffeo's collaborative agent dynamically queries partner data on behalf of users
Internal Users
Manual query process
Diffeo partner portal
Diffeo partner data

Figure 8: Diffeo integration with third-party data providers.

Diffeo has a relationship with Reed Elsevier's LexisNexis business unit and has integrated three of their subscription data sources. Currently, the LN Moreover news feed is activated in `diffeo.com`, and we are looking for customers interested in activating other LN data sources.

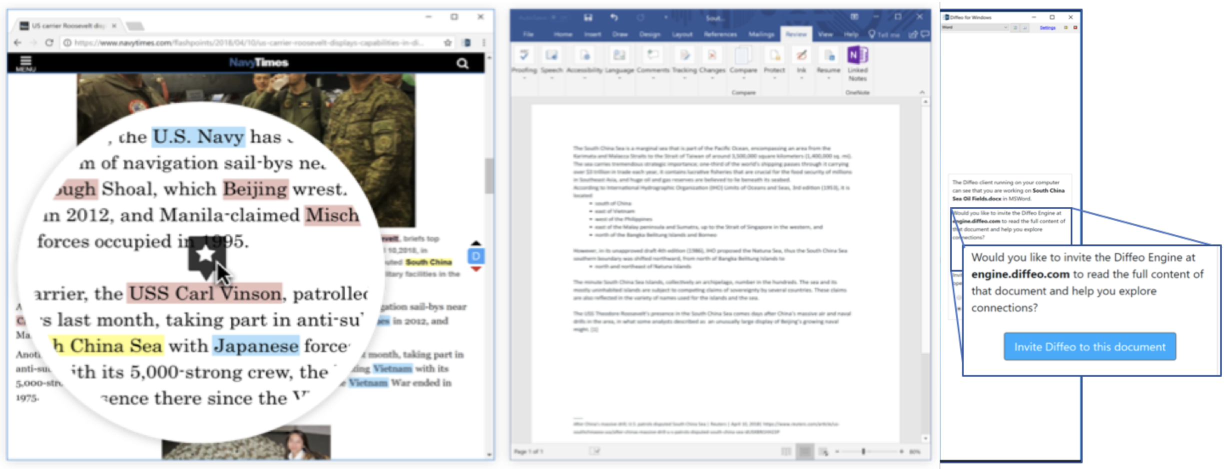
3. Algorithms

3.1. User-Facing Sensors

As shown in Figure 9, Diffeo integrates directly with user's familiar tools.

The Diffeo for Browsers (DfB) extension installs in Chrome and Firefox, and inserts highlight-based recommendations in PDFs and web pages. The star button shown in the tooltip over the mention of USS Carl Vinson enables users to indicate interest in a recommended entity, which adds the entity to the project journal.

The Diffeo for Windows (DfW) desktop client installs on the Windows desktop and communicates with Microsoft Office tools via COM APIs. As shown on the rightmost portion of Figure 9, users explicitly invite Diffeo into documents.


NavyTimes
, the U.S. Navy has
m of navigation sail-bys ne
ugh Shoal, which Beijing wrest.
n 2012, and Manila-claimed Misch
forces occupied in 1995.
arrier, the USS Carl Vinson, patrolle
's last month, taking part in anti-su
h China Sea with Japanese force
ith its 5,000-strong crew, the
ence there since the V
Diffeo for Windows
Invite Diffeo to this document

Would you like to invite the Diffeo Engine at **engine.diffeo.com** to read the full content of that document and help you explore connections?

Figure 9: visual affordances for collaboration between humans and machines.

3.2. Operational Transforms Journals

To track the context in a user's current project, the Diffeo system maintains an append-only log of knowledge operations, such as starring a mention of interest or presenting evidence of a relation. As illustrated in Figure 10, this synchronizes the Diffeo interactions across the many document views into which a user might invite Diffeo.

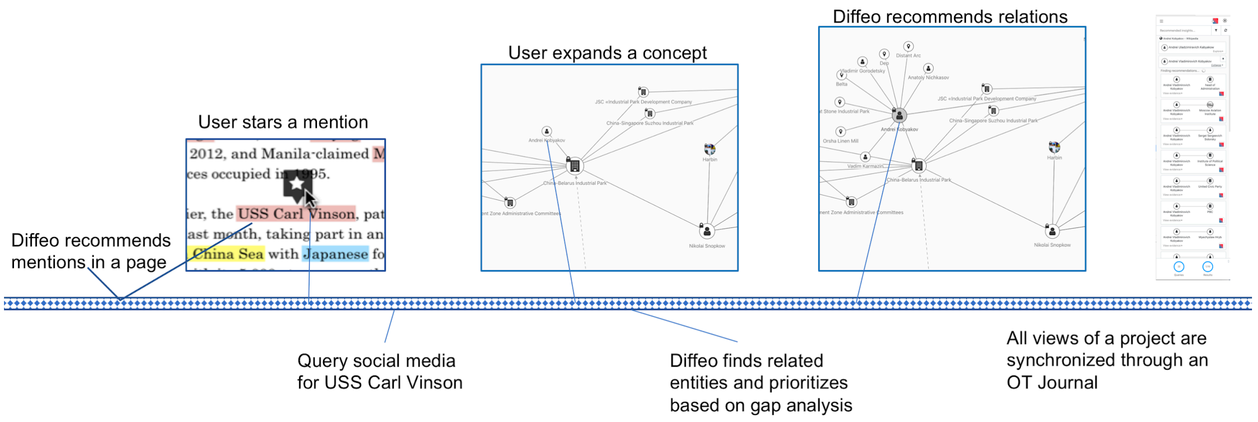

Diffeo recommends relations
User expands a concept
User stars a mention
Diffeo recommends mentions in a page
2012, and Manila-claimed M
ces occupied in 1995.
ier, the USS Carl Vinson, pat
ast month, taking part in an
China Sea with Japanese fo
Andrei Kobyakov
China-Belarus Industrial Park
China-Singapore Suzhou Industrial Park
JSC «Industrial Park Development Company
Harbin
Nikolai Snopkow
Query social media for USS Carl Vinson
Diffeo finds related entities and prioritizes based on gap analysis
All views of a project are synchronized through an OT Journal

Figure 10: Diffeo knowledge operations journals

3.3. Bigforest

Human communication, especially language, is inherently ambiguous. Interpreting communications requires judgment, and such judgments involve confidence assessments. Dr. Onyshkevych points out that unstructured communication entails at least seven forms of confidence—or lack thereof! These include quantifying the possibility of deception and omitted information. In Diffeo, we generally focus on interpreting the intent of the author.

Most communications can be broken into smaller parts, such as textual phrases, audio utterances, and subimages. A communicator's intent is conveyed through the intent of these smaller parts. Component-level tasks in natural language processing (NLP), speech processing, and image analysis involve resolving the meaning or intent of these smaller parts.

Diffeo's forest graph environment is called "Bigforest" and is illustrated in Figure 12. As discussed in Section 2, humans resolve ambiguous communication by identifying common references. Usually, a common reference is itself another communication element, such as a Wikipedia article or entry in the OED. Such intent resolution is part of learning to communicate in a language. The term "co-reference" or "coref" refers to this process of understanding through identifying shared references. In Diffeo slang, one can say that a set of coreferent data elements is a set of corefs.

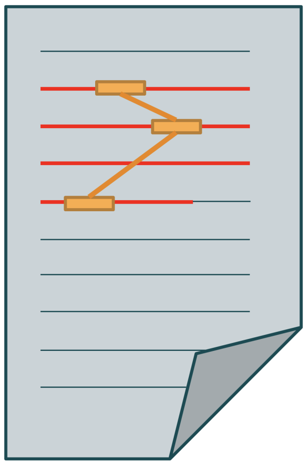


Figure 11: mention chain in a document

Automatic systems can also attempt to resolve meaning. This document delineates types of disambiguation algorithms. Unlike essentially all other entity-centric text processing systems, the Diffeo system does **not** focus on finding **all** mentions of an entity or concept. Quite the opposite, the Diffeo system focuses on finding the *surprising* coreferences.

As a person gains context on a subject, s/he gathers reference for communicating with others and his/her self. This set of known coreferences defines a retrieval target for the subject. A user expands this set by identifying its boundary and finding connections that go beyond it. The Diffeo disambiguation system seeks to accelerate this discovery process by organizing the mentions for exploration.

Diffeo's indexing system builds three tiers of indices. Tier0 is traditional full-text keyword and phrase indexing. Tier1 enriches the data records using state-of-the-art natural language processing, including the Rosette Linguistics Platform (RLP) from Basis Technology, which is included in the Diffeo platform. Tier1 enables smart tags and feature tensors. The Bigforest hierarchical coreference system powers Tier2 indexing, which groups mention by probability of coreference and enables relation discovery by bundling evidence of connections. This multi-tier indexing system scales horizontally using Kubernetes to support massive data.

As depicted in Figure 11, Diffeo's Tier1 pipeline uses advanced natural language processing models to recognize mentions of concepts and entities. This process defines semantically meaningful units within a document by grouping mentions of the same thing together into "coreference chains" within each document. The Diffeo system refers to these objects as "MentionChains" and they are separately indexed within Elasticsearch. The access control lists and other metadata properties from the parent document are propagated into the MentionChain objects.

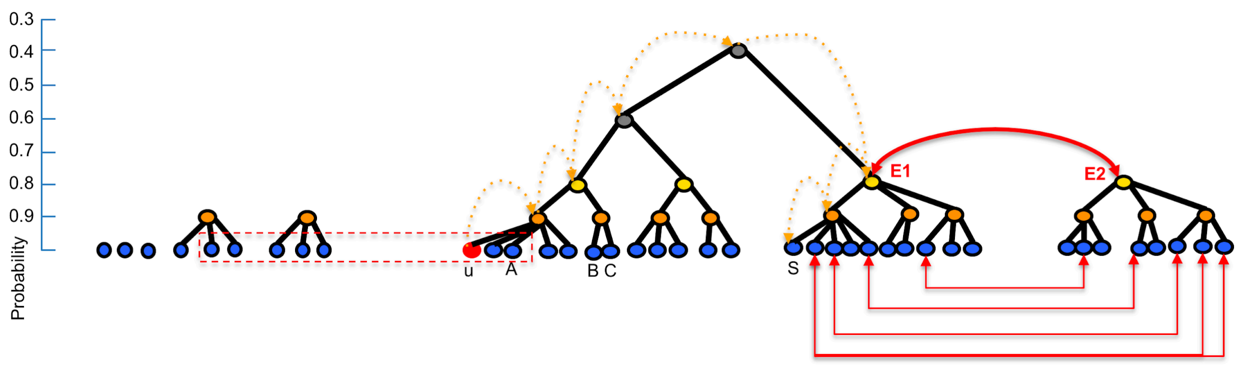


Figure 12: Bigforest maintains nested sets that encoded likelihood of coreference. See also the Bigforest Infographic in the appendix.

Diffeo's hierarchical agglomerative clustering framework provides state-of-the-art entity disambiguation performance at massive scale. This framework digests natural language data from structured and unstructured sources to create feature tensors, which form the leaf-level vertices in a large forest graph.

3.3.1. Bigforest Data Flow

1. The forest of tree graphs lives in **probability space.** Each leaf vertex represents natural language mentions of an entity from an indexed document or a knowledge base article about an entity. Lower interior vertexes connect leafs that are closer together, i.e. more likely to be the same entity (aka "coreferent").

2. Interior vertices carry **summaries** of context data from within-doc coref chains of mentions represented by leafs below.

3. Incremental batch clustering selects blocks of mentions based on feature data queries (see red dashed box).

3.1 Blocking applies pairwise feature vector comparisons to high-recall subsets of all possible pairs of leafs.

3.2 Overlap regions of blocks require merging.

4. Mentions are clustered together based on probability of coreference, which captures transitive closure recall gains, A=B U B=C ➔ A=C even if A and C appeared distant. The system uses the 70% probability level as the mention level.

5. Subtrees below the entity level represent subtopics within an entity. The induced structure enables fast novelty ranking.

6. Roots above entity level connect confusable entities enabling a profile leaf (P) to efficiently discover a surprising mention (S), see orange dotted arc in Figure 2.

7. Resolved co-occurring mentions enable open-domain cross-document relationship resolution. Red lines represent passages in source documents that mention both Entity 1 and Entity 2.

8. Summary vertexes aggregate **entity-resolved relations** for ranking and latent knowledge graph analytics.

3.3.2. Order Complexity of Entity Disambiguation

Intent resolution can be simple. Given input data from a known domain and a list of strings with unique meanings in that domain, a simple string match algorithm is linear in the input size.

General intent resolution can be expressed as a partitioning problem: given a set of elements from an input corpus, such as a set of mentions, generate a complete partitioning of the elements into coreferent sets. The number of possible partitionings of N objects is Bell(N) is large, so if a particular algorithm must evaluate all of them, then it is expensive.

N could be documents or mentions or mention chains. A typical enterprises have tens of millions to billions of documents. Each document has tens to hundreds of mention chains.

Grappling with such expensive order complexities has become an obsession among some computer scientists. It turns out that such glorious algorithmic challenges are quite unnecessary from the perspective of end users—they are a distraction. Users actually care about much less than the entire partitioning, and optimizing for the entire partitioning steers away from the discovery challenge actually faced by users. Before returning to the discovery challenge below, let's discuss some of the approaches to coping with the order complexity of coref.

Diffeo has experience in working on large scale meaning resolution. In 2011, we started developing a clone of the hierarchical disambiguation algorithms explored by Andrew McCallum's group. See, for example, the Markov Chain Monte Carlo (MCMC) hiercoref prototype written in Scala in FactorIE developed by Michael Wick for A Discriminative Hierarchical Model for Fast Coreference at Large Scale. That algorithm was motivated by an earlier paper by Sameer Singh: Large-scale cross-document coreference using distributed inference and hierarchical models, which explored the efficiency of using MCMC[4] as a technique for clustering mentions from a large corpus. The exciting premise was that the coref problem could be parallelized across workers that evaluate individual "moves" that assemble a partition iteratively. Instead of directly evaluating all possible partitions, which is the staggeringly large number Bell(N)[5], a system can iteratively hill climb toward the best partition by considering individual small edits to an existing partition.

The specific mechanism for evaluating an individual move is the Monte Carlo criterion from physics: given an "energy" function, a candidate move is accepted if it lowers the energy ***or*** if it increases the energy by an amount that is less than a randomly selected amount of noise. This noise enables the system to (eventually) escape local minima in the energy function, such that the system can (eventually) discover the lowest energy configuration. For this approach to work, one must design an energy-like function for the data at hand, which means a function that generates a similarity score between mentions.

[4] See also the well regarded: The MCMC Revolution by Diaconis
[5] https://en.wikipedia.org/wiki/Bell_number

Hierarchical clustering is useful for coreference. To understand why, we must look at how the energy-like function used in MCMC for natural language differs from the energy functions used in physics. Physical energy functions model situations that generally involve real-valued, continuous functions. In contrast, **natural language data is discrete and sparse**. As a result, the pseudo-energy functions used for inference in natural language often output "none" or **"no info."** This does not happen in a physical model.

For example, consider a similarity function implemented by comparing words that appear in the sentences surrounding a textual mention. Even for two coreferent phrases, they might have no context words in common. In fact, they *often* won't. As a result, the word comparison yields a nil output, so there is no information for estimating the probability of coreference. The likelihood function in a typical model therefore simply outputs zero. When combined with a Bayesian prior, such a model will output the prior probability, i.e. the background probability that two randomly selected mentions are coreferent, e.g., for a coin toss, the prior is 0.5[6].

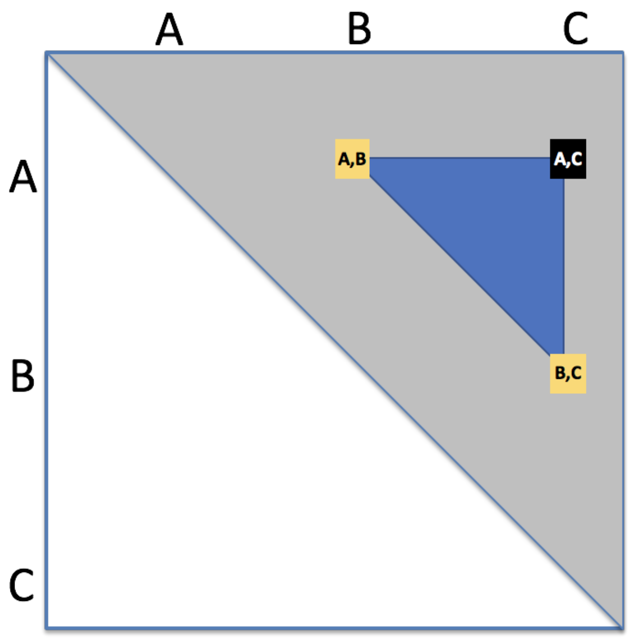


Figure 13: A pairwise matrix describing coref probabilities is highly constrained by transitive closures. For example, if we know that A=B and that B=C, then we also know that A=C, even if a pairwise model lacked sufficient information to be confident about comparing A and C directly. Thus, missing cells can be filled using other values.

There are two general approaches to coping with this problem of **no info**:

1. Project the context features into **lower dimensional spaces** that have a better chance of overlap without introducing too much noise.
2. Exploit gains from ***transitive closures (Figure 13)***. If Chairman Smith is John Smith Jr., and Chairman Smith is driving the boat, then John Smith Jr. is driving the boat. Abstractly, this is the mathematical notion of transitivity: if A=B, and A=C, then B=C. Despite the lack of overlapping contexts between particular pairs of coreferent phrases, humans can still understand them by identifying a common coreference—we intuitively use transitivity every day. Real coreference algorithms must cope with varying degrees of confidence in coreference assertions, see below about cophenetic distance.

These two approaches can be visualized by considering the pairwise matrix of similarities between data elements. The energy function used in Monte Carlo is constructed by considering whether a particular change in the partition would group together pairs of data elements with greater similarity. The sparseness problem is manifested in empty cells in the pairwise matrix, i.e. cells for which no information exists. Feature densification improves this, i.e. it turns on more

[6] In practical coref systems, the prior is usually intimately tied to the canopy or blocking function that generates candidates for a limited budget of coref comparisons.

of the cells. Transitivity detects triples of cells arranged in right triangles in which at least two of the three cells are on, see Figure 13.

Hierarchical agglomerative clustering is a general approach for both densifying features and capturing transitive gains. Specifically, the agglomeration step introduces new interior vertices that appear in the pairwise matrix as new rows. These new vertices carry agglomerated features from the leaf-level vertices below them, and these new vectors have greater overlap — i.e. are denser. As representatives of a set of coreferent elements, these interior vertices can connect subsets of coreferent elements that individually share little or no context with each other.

Hierarchical clustering also directly addresses ambiguity in transitive closures. The dendrograms produced by clustering algorithms like linkage (see fastcluster[7]) estimate the cophenetic distance between interior vertices. We use the `average` linkage function as the best approximation for transitive closures.

As illustrated in the Singh paper, it can take a very long time for MCMC algorithms to refine a tree to even the partition accuracy of name-match. Fortunately, MCMC is not the only way to build hierarchical agglomerative clusters. The reason MCMC takes so long is that it gets stuck in critical slowing down, which makes it exponentially slow. It gets stuck exploring all possible worlds rather than searching directly for the most probable worlds. We explored this by making a set of MCMC moves for hierarchical clustering that obey detailed balance. This helped us arrive at the conclusion that that there is no move like the famous Swendson-Wang move for hierarchical clustering.

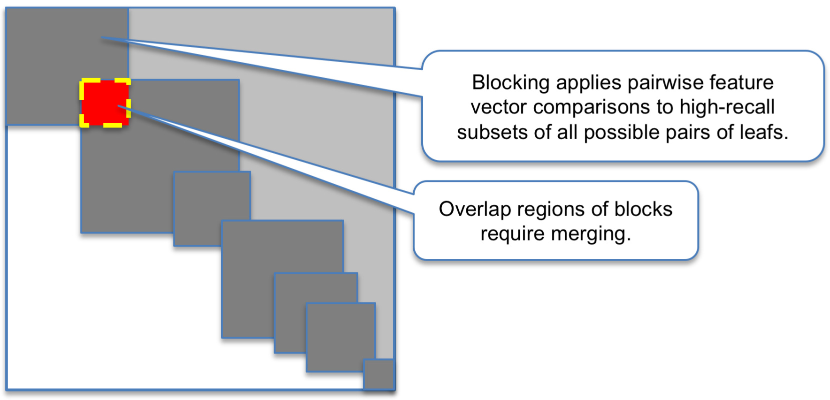

Blocking applies pairwise feature vector comparisons to high-recall subsets of all possible pairs of leafs.
Overlap regions of blocks require merging.

Figure 14: Canopy batches down the diagonal of the global pairwise matrix.

There are many approaches to hierarchical agglomerative clustering. In particular, the family of linkage based clustering algorithms are efficient algorithm for generating clusters that optimize a quality function. Linkage clustering is typically executed in a greedy fashion, which makes it unable to handle streaming data sets or data sets that do not fit in memory. While MCMC is slow, it does handle large data sets that change over time. Generally, linkage algorithms are run on the entire pairwise matrix of dissimilarities between data elements. The linkage algorithm efficiently finds the best groupings based on distances between the leaf-level items.

To address this, Diffeo's Bigforest framework uses an ***iterative*** linkage algorithm that we call plink for "parallel linkage." Plink runs linkage in batches, and merges the resulting trees. The merge algorithm relies on interior vertices acting as good summaries of their leaf-level vertices that represent mentions. One can picture this as overlapping blocks going down the diagonal of the pairwise matrix, see Figure 14.

[7] http://danifold.net/fastcluster.html

Linkage can operate in O(n^2), and the asymptotic order complexity of batched linkage is the same. However, the prefactor matters, and such blocking or canopying has a big impact on practical implementations of large scale clustering.

Diffeo uses the Kodama library[8] for linkage, which is implemented in Rust and offers an interface for Golang.

3.3.3. Semi-Structured Entity Records

Another technique for significantly reducing the prefactor of large-scale coreference resolution is high-precision linking, which we call wikification. In Diffeo’s internal jargon, we reserve the word “wikification” to mean a disambiguation algorithms that operates on a single document and output a high-precision set of mention resolutions in the form of references to external knowledge bases, like Wikipedia. This can be visualized as an attempt to eliminate some of the largest blocks in the block diagonal view of the pairwise matrix. It relies on the external world knowledge providing enough context data that high-precision coreference assertions can be made.

Running all of the possible canopies for plink is large, so we prioritize the clustering work units using delta name count statistics. As data streams in, Tier1 ingest increments name counts that determine which clustering jobs get created. When a job gets run on a set of mentions, the associated name count gets decremented.

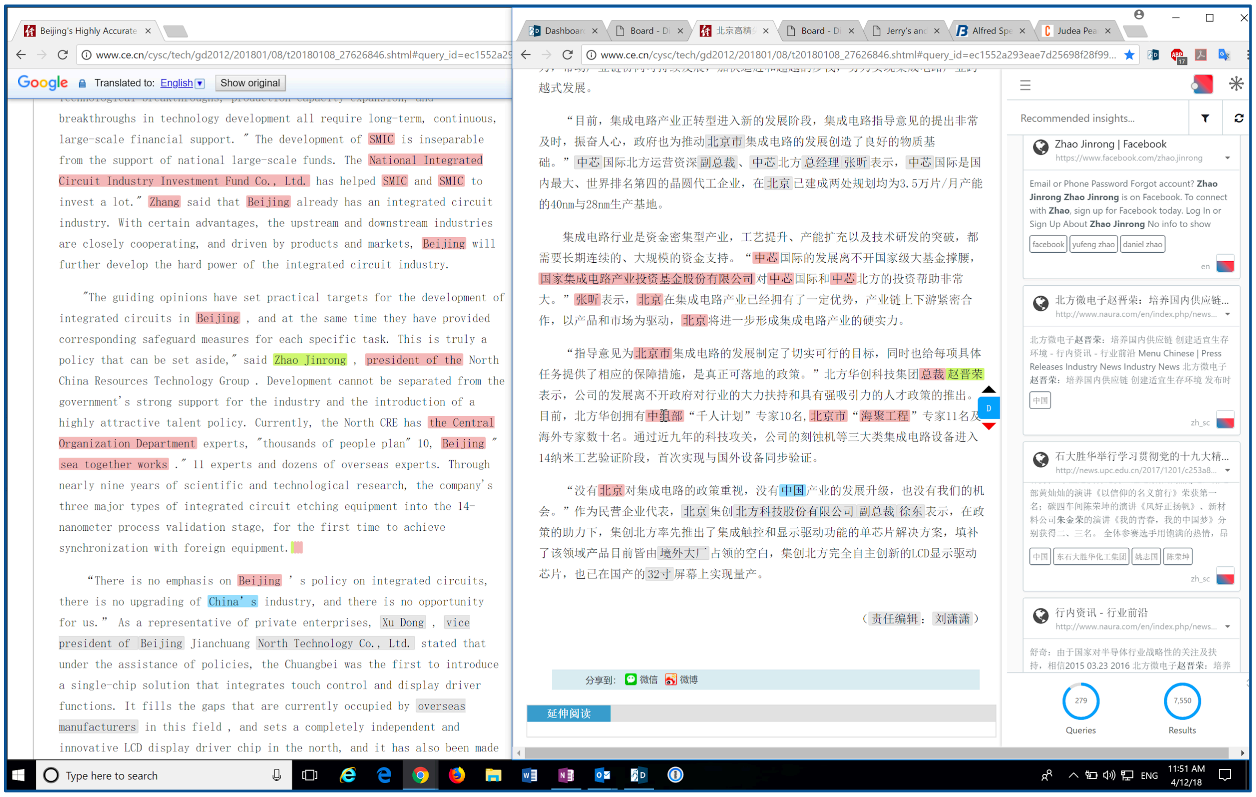
The images and icons shown in Figure 16 come from wikification to well known reference sources, such as Wikipedia and Twitter.

To incorporate structured data, Diffeo offers an API for loading external entity profile records. An external service can enumerate a set of known entities and assign identifiers to them. An entity record is the detailed structured information available from a single external source about a single entity, distinguished on that source's identifier; for instance, a single Facebook user profile. This may be made up of several **raw entity subrecords**, such as the profile data and connected entities. Loading entity record data is a simple task of mapping the external field names into Diffeo suit of extractors.

3.3.4. Cross-Lingual Entity Discovery

Diffeo uses several translation and transliteration technologies in order to find content in other languages. The cross-lingual process starts with formulating queries in other languages, such as Chinese and Russian, and sending those queries to appropriate search engines via the Metasearch infrastructure. The pairwise coreference model uses an extensive set of cross-lingual features for entity disambiguation, so that it can generate examples like this Figure 15.

[8] https://github.com/diffeo/kodama


Beijing's Highly Accurate
www.ce.cn/cysc/tech/gd2012/201801/08/t20180108_27626846.shtml#query_id=ec1552a29
Google Translated to: English Show original
Dashboard
Board - Di
北京高精
Board - Di
Jerry's and
Alfred Spe
Judea Pea
www.ce.cn/cysc/tech/gd2012/201801/08/t20180108_27626846.shtml#query_id=ec1552a293eae7d25698f28f99...
Recommended insights...
Type here to search
ENG
11:51 AM
4/12/18

breakthroughs in technology development all require long-term, continuous, large-scale financial support. ″ The development of SMIC is inseparable from the support of national large-scale funds. The National Integrated Circuit Industry Investment Fund Co., Ltd. has helped SMIC and SMIC to invest a lot.″ Zhang said that Beijing already has an integrated circuit industry. With certain advantages, the upstream and downstream industries are closely cooperating, and driven by products and markets, Beijing will further develop the hard power of the integrated circuit industry.

″The guiding opinions have set practical targets for the development of integrated circuits in Beijing , and at the same time they have provided corresponding safeguard measures for each specific task. This is truly a policy that can be set aside,″ said Zhao Jinrong , president of the North China Resources Technology Group . Development cannot be separated from the government's strong support for the industry and the introduction of a highly attractive talent policy. Currently, the North CRE has the Central Organization Department experts, ″thousands of people plan″ 10, Beijing ″ sea together works .″ 11 experts and dozens of overseas experts. Through nearly nine years of scientific and technological research, the company's three major types of integrated circuit etching equipment into the 14-nanometer process validation stage, for the first time to achieve synchronization with foreign equipment.

"There is no emphasis on Beijing ' s policy on integrated circuits, there is no upgrading of China' s industry, and there is no opportunity for us." As a representative of private enterprises, Xu Dong , vice president of Beijing Jianchuang North Technology Co., Ltd. stated that under the assistance of policies, the Chuangbei was the first to introduce a single-chip solution that integrates touch control and display driver functions. It fills the gaps that are currently occupied by overseas manufacturers in this field , and sets a completely independent and innovative LCD display driver chip in the north, and it has also been made

越式发展。

“目前，集成电路产业正转型进入新的发展阶段，集成电路指导意见的提出非常及时，振奋人心，政府也为推动北京市集成电路的发展创造了良好的物质基础。”中芯国际北方运营资深副总裁、中芯北方总经理张昕表示，中芯国际是国内最大、世界排名第四的晶圆代工企业，在北京已建成两处规划均为3.5万片/月产能的40nm与28nm生产基地。

集成电路行业是资金密集型产业，工艺提升、产能扩充以及技术研发的突破，都需要长期连续的、大规模的资金支持。“中芯国际的发展离不开国家级大基金撑腰，国家集成电路产业投资基金股份有限公司对中芯国际和中芯北方的投资帮助非常大。”张昕表示，北京在集成电路产业已经拥有了一定优势，产业链上下游紧密合作，以产品和市场为驱动，北京将进一步形成集成电路产业的硬实力。

“指导意见为北京市集成电路的发展制定了切实可行的目标，同时也给每项具体任务提供了相应的保障措施，是真正可落地的政策。”北方华创科技集团总裁赵晋荣表示，公司的发展离不开政府对行业的大力扶持和具有强吸引力的人才政策的推出。目前，北方华创拥有中组部“千人计划”专家10名，北京市“海聚工程”专家11名及海外专家数十名。通过近九年的科技攻关，公司的刻蚀机等三大类集成电路设备进入14纳米工艺验证阶段，首次实现与国外设备同步验证。

“没有北京对集成电路的政策重视，没有中国产业的发展升级，也没有我们的机会。”作为民营企业代表，北京集创北方科技股份有限公司副总裁徐东表示，在政策的助力下，集创北方率先推出了集成触控和显示驱动功能的单芯片解决方案，填补了该领域产品目前皆由境外大厂占领的空白，集创北方完全自主创新的LCD显示驱动芯片，也已在国产的32寸屏幕上实现量产。

（责任编辑：刘潇潇）

分享到： 微信 微博

延伸阅读

Zhao Jinrong | Facebook
https://www.facebook.com/zhao.jinrong
Email or Phone Password Forgot account? **Zhao Jinrong Zhao Jinrong** is on Facebook. To connect with **Zhao**, sign up for Facebook today. Log In or Sign Up About **Zhao Jinrong** No info to show
facebook | yufeng zhao | daniel zhao
en

北方微电子赵晋荣：培养国内供应链...
http://www.naura.com/en/index.php/news...
北方微电子**赵晋荣**：培养国内供应链 创建适宜生存环境 - 行内资讯 - 行业前沿 Menu Chinese | Press Releases Industry News Industry News 北方微电子**赵晋荣**：培养国内供应链 创建适宜生存环境 发布时
中国
zh_sc

石大胜华举行学习贯彻党的十九大精...
http://news.upc.edu.cn/2017/1201/c253a8...
部黄灿灿的演讲《以信仰的名义前行》荣获第一名；碳四车间陈荣坤的演讲《风好正扬帆》、新材料公司**朱金荣**的演讲《我的青春，我的中国梦》分别获得二、三名。 全体参赛选手用饱满的热情，昂
中国 | 东石大胜华化工集团 | 姚志国 | 陈荣坤
zh_sc

行内资讯 - 行业前沿
http://www.naura.com/en/index.php/news...
舒奇：由于国家对半导体行业战略性的关注及扶持，相信2015 03.23 2016 北方微电子**赵晋荣**：培养

279 Queries
7,550 Results

Figure 15: example of Diffeo discovering a mention of an entity in a different language from the language that the user first indicated interest. In this example, the user's working notes referred to Zhao Jinrong in English, and Diffeo found an interesting mention of him in Chinese. The left-most window shows this page translated into English to show that the yellow highlighted phrase is the target entity.

3.3.5. Coref Metrics

Measuring the accuracy of coref algorithms is an ongoing area of research interest for many academic researchers. Interestingly, name-match based coref gets an F_1 score in the 50-60% range on many real-world problems. Even for the famous John Smith data set, in which all of the entities have the name John Smith, the name-match coref score gets an F_1 score in the mid-fifty percent range, because one of the larger clusters is "John Smith Jr."

Two other baseline systems for coref are shatter-all and cluster-all, which correspond to assigning every input element to a distinct cluster or to one giant cluster, respectively. For any given data set, the triad of scores for these three baseline algorithms goes a long way toward characterizing what is easy and hard about the data set.

F_1 is a basic metric used for evaluating binary classifiers. To apply it to coref, one must map the partitioning problem into a binary assignment problem. There are several prescriptions for this mapping. For many years, the most widely used mapping was BBB F_1 put forward by Bagga and Breck Baldwin in Algorithms for Scoring Coreference Chains (1998). More recently Constrained Entity-Alignment F-Measure (CEAF) has gained acceptance, because it is conceptually more appealing and gives lower scores to cluster-all and shatter-all. One can compute BBB F_1 for shatter-all and cluster-all using the frequency distribution of correct partition sizes see Diffeo Technical Note #5.

F_1 is a set-based metric. IR evaluations usually use rank-based measures, see Section 1.3.

3.5 Superworkers

To process lots of data, a standard architectural pattern is to break the processing into steps that can be scaled horizontally using different resources for each step. This enables massive scaling and burstability. It also enables decoupled implementation. For example, step one can be a third-party thing written in Java, step two can be CPU optimized native binary compiled from Golang, and step three can be a still-in-research prototype written in python.

However, that standard approach fails when the underlying task is so massive that one should not expect to ever be able to do *all of the possible work*. For such tasks, other architectures enable intelligent prioritization of the *most useful work*. Learning and analysis tasks are examples of such unbounded tasks. Given a finite budget of time and resources, what are the most productive relationships and entities to explore?

Instead of breaking the problem into small pieces and requiring that most of the subtasks be finished quickly, Diffeo uses the superworker design pattern to concentrate the available information in a prioritization loop that cuts combinatorially large problems, like Figure 16a, down to budgeted problems, like Figure 16b, so that it can find the most useful and interesting subgraphs, like Figure 16c.

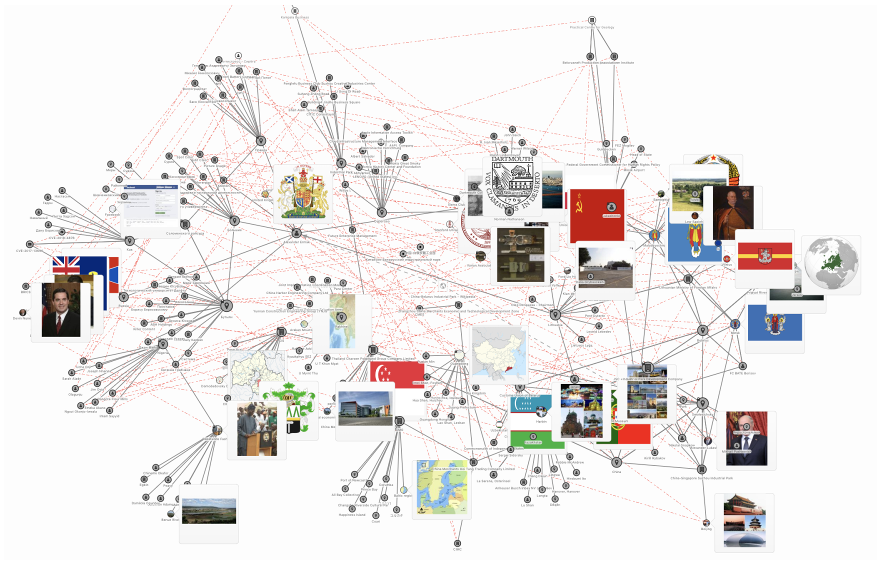


Figure 16a: the number of possible subgraphs increases in a combinatorial explosion.

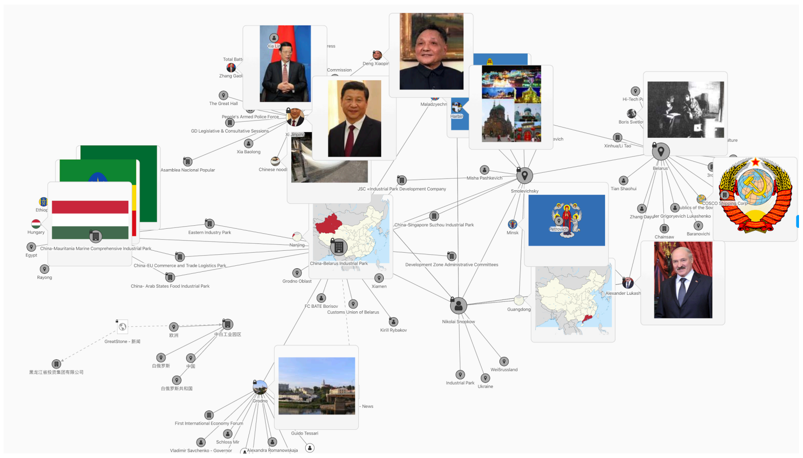
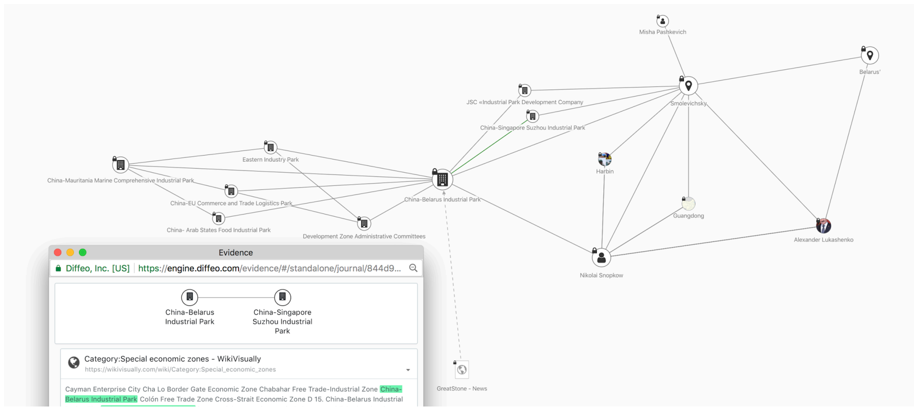


Figure 16b: intelligent prioritization can efficiently explore high value connections within a time budget with limited compute resource.


Misha Pashkevich
Belarus'
Smolevichsky
JSC «Industrial Park Development Company
China-Singapore Suzhou Industrial Park
Eastern Industry Park
China-Mauritania Marine Comprehensive Industrial Park
China-EU Commerce and Trade Logistics Park
China- Arab States Food Industrial Park
Development Zone Administrative Committees
China-Belarus Industrial Park
Harbin
Guangdong
Alexander Lukashenko
Nikolai Snopkow
GreatStone - News

Evidence

Diffeo, Inc. [US] | https://engine.diffeo.com/evidence/#/standalone/journal/844d9...

China-Belarus Industrial Park — China-Singapore Suzhou Industrial Park

Category:Special economic zones - WikiVisually
https://wikivisually.com/wiki/Category:Special_economic_zones

Cayman Enterprise City Cha Lo Border Gate Economic Zone Chabahar Free Trade-Industrial Zone China-Belarus Industrial Park Colón Free Trade Zone Cross-Strait Economic Zone D 15. China-Belarus Industrial

Figure 16c: such efficient exploration enables Diffeo to find useful recommendations.

Appendices

Vocabulary

Diffeo's software represents knowledge as mentioned **concepts** and **relations** between concepts. Diffeo focuses on **relationships between entities**, which are a special type of concept. The following vocabulary words are useful for discussing how Diffeo works.

A ***concept*** is a mentionable idea. Mentioning includes uttering, depicting, and writing. A concept is only defined in the context of communication between actors.

- Contrast with ***noncept***, i.e. an idea for which no atomic phrase has been coined.
- A concept is defined by its mentions.
- A concept has no existence other than its mentions, which includes utterances, depictions, and scribings. That is, a concept is evidenced only by its mentions.
- We no access to truth, only evidence. We can only infer truth from evidence.

An **event** is a concept ***defined,*** at least in part, ***by*** its spacetime location.

An **entity** is an instance of a ***class of events*** that are each distinguished by strongly typed attributes, such as hometown, phone number, number of stories, DNA.

- Named entities have a given name, generally not unique.
- Pronominal mentions of entities refer to an antecedent. E.g., him, you.
- Nominal mentions of entities describe a particular entity without providing a name, e.g. “the general” or “the three female journalists.”
- Examples of **entity classes**:
    - Person
    - Company or Organization
    - Facility or Building
    - Vehicle
    - Device or cyber identifier, e.g. phone number, skype handle, email address, IP address
    - Chemical Compounds
    - Planetary Objects
    - Proteins

An **actor** is an entity present in the current environment, e.g. a person in the room right now. For digital environments, examples include a user or an email correspondent or a voice on the phone.

- Distinguished from non-actor entities by having active properties or methods that can be invoked with an input to receive a response.
- Many inanimate objects can be instances of a class of entities and still not qualify as actors. For example, Oxygen is an instance of the the entity class called Chemical Elements. However, Oxygen itself is not an actor, and an individual Oxygen atom or O_2 molecule is often not ***distinguishable*** from other Oxygen atoms and thus not even an entity.

A **document** is a digital media object containing mentions of concepts.

- Non-digital records, such as the printed form of a Document is merely an artifact and not a Document in our sense of the word.
- Documents appear in many contexts. Two contexts of particular importance are:
    - Source documents that a user might use as reference material, and
    - Working notes that a user is creating, such as a reply in an email thread.

A **relationship** is a kind of attribute that entities have.

- A relationship is identified by two entities, which are concepts, and therefore substantiated by mentions — at least one mention for each of the two entities.
- A relationship can be described as having a “type” that is defined by a document, often called an ontology. The mentions of various concepts in an ontology, such as “member of” are used to describe relationships between entities, such as “Black Francis was a

member of The Pixies." The two-tuple of (Black Francis, The Pixies) is a more fundamental object.

Evidence bundles consist of documents that substantiate a relationship between two events.

Coreference resolution or **coref:** Disambiguation is the process of recognizing which meaning a mention is intended to have. This is subtle, because nobody has direct access to meaning or truth. Instead, we can only infer truth from evidence. Thus, the literal meaning of the word disambiguation is actually unobtainable. Instead, the only operationalizable notion of disambiguation is coreference resolution. If an actor believes that a particular mention is referring to the same concept as another mention, then the actor believes that the two mentions are co-referring or co-referent, i.e. have the same meaning. One says "I understand X" or "I know the meaning of X" when one has associated X with other concepts through coreference resolution and awareness of some of its relations. To connect a mention to other mentions of the same concept is to ***resolve*** its meaning, or to "coref" it, after which it has been "**coref'ed**."

- **Blocking:** is a design pattern that selects a coarse set of partitions using a shallow coref algorithm so that more expensive deep coref algorithms can consider each block separately rather than all of the data at once. By definition, blocks are non-overlapping.
- **Canopying:** is like blocking except with overlapping selections. In general, canopies allow better recall, because blocking prevents considering alternatives that cross blocking divisions.
- **Deep Coref Algorithm:** in addition to using all the available data on individual data elements, a complete coref algorithm also exploits transitive closures and densified features. In the ideal limit, a "complete" coref algorithm would use all of the data that a human might use to detect coreferences.
- **Shallow Coref Algorithm:** an algorithm that generates sets of coreferent elements using only some of the available data, such as name match or partial name match.

Coref of an agents carries extra weight because it identifies the actor. Identity has many practical implications for actors in the real world. Since many online cyber actors cloak their identities, we use the special term **persona** to refer to an online actor that another actor may not yet have accurately coref'ed to other personas.

A **context** is a portion of a document that enables or helps a person to understand the intended communication of mentions within that context. Context is variable. The boundaries of context are ill defined. Generally, including a larger portion as context will enable the reader to understand more. On the other hand, larger context requires more time to digest. For example, a snippet of text in a search result listing is intended to give the reader enough context to efficiently decide whether or not to open the link to access the full document.

A **citation** is a reference to a document. A URL is a kind of citation. However, a more complete citation provides enough information for a reader to disambiguate a specific version of a document, whereas a URL usually does not. When a document contains a citation to another document, the **context** of the citation usually indicates some concepts that are mentioned in both.

Utility Function or Goal

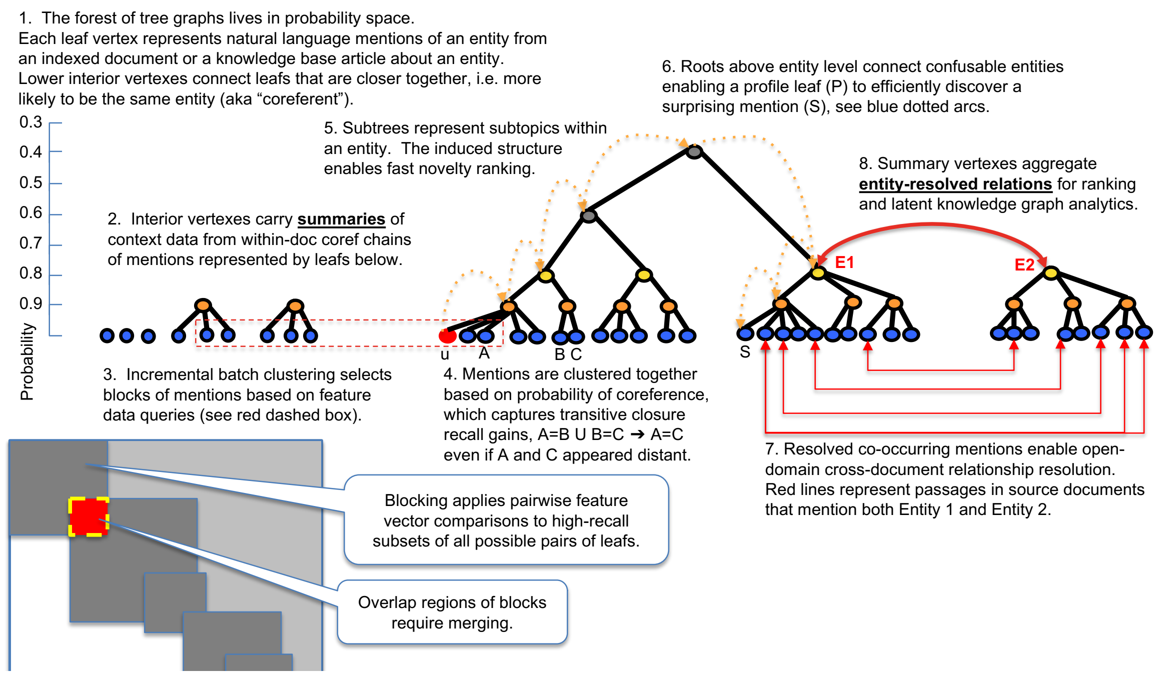
- Actors are often called “agents” because they have agency, which is to say the behavior of its active methods is accurately described by ascribing ***goals*** to the actor.
- An actor’s utility function is a description of those goals, and is often reduced to logic or math in order to enable computation at the expense of accuracy. That is, we reduce the fidelity of the utility function in order to make it accessible to machines.

Learning is a process of accumulating knowledge in pursuit of optimizing a utility function. Amazingly, the learning process can change the goal! However, that is out of scope for this document.

In scope for this document is the idea that a person (**actor**) can seek to **learn** about a ***relationship*** by obtaining **evidence** about it. Further, a useful source of such evidence is often other actors in the environment. The Diffeo recommender engine is an actor that learns from human actors.

The Diffeo system is a software-based actor that has an objective function defined by a user’s current working context. Specifically, the Diffeo actor’s objective function is to find evidence of relations between entities mentioned in the user’s working context and other entities that are not yet in the user’s working context such that the user adds (cites) those documents to the working context. The specific implementations of this objective function within the current Diffeo system mimic human processes of querying remote systems, retrieving documents, digesting content, corefing entities, and looking to fill gaps between a human collaborator’s working context.

Bigforest Infographic


1. The forest of tree graphs lives in probability space. Each leaf vertex represents natural language mentions of an entity from an indexed document or a knowledge base article about an entity. Lower interior vertexes connect leafs that are closer together, i.e. more likely to be the same entity (aka “coreferent”).

2. Interior vertexes carry **summaries** of context data from within-doc coref chains of mentions represented by leafs below.

3. Incremental batch clustering selects blocks of mentions based on feature data queries (see red dashed box).

4. Mentions are clustered together based on probability of coreference, which captures transitive closure recall gains, A=B U B=C ➔ A=C even if A and C appeared distant.

5. Subtrees represent subtopics within an entity. The induced structure enables fast novelty ranking.

6. Roots above entity level connect confusable entities enabling a profile leaf (P) to efficiently discover a surprising mention (S), see blue dotted arcs.

7. Resolved co-occurring mentions enable open-domain cross-document relationship resolution. Red lines represent passages in source documents that mention both Entity 1 and Entity 2.

8. Summary vertexes aggregate **entity-resolved relations** for ranking and latent knowledge graph analytics.

Blocking applies pairwise feature vector comparisons to high-recall subsets of all possible pairs of leafs.

Overlap regions of blocks require merging.