Hierarchical document classification: simplifying search

wipro

Organizations are sinking in a deluge of documents. There are mountains of documents dealing with employee records, customers, equipment, production data, research, trade, financial records, stock markets, news, education, entertainment, governance; the list is endless. Library science emerged to classify these documents for quick and easy retrieval. But traditional classification methodologies, limited by their manual nature, are unable to keep pace with the growth in documents. If that isn't bad enough, there is a secondary problem. Along with the growth in documents has come a growth in the number of categories. Today, every domain has broader and deeper specialization. The impact of these changes is becoming apparent: if a document is poorly classified, it may never be found resulting in a business loss which business team can't afford.

Let's put the problem in perspective. Think of a sportscaster looking for data on a cricket match. A report on a cricket match can be classified in too many ways: using year, teams, players, scores, commentators reporters/writers and many more. The person tagging such a report needs to be an expert at the game without which the report (document) may never get tagged properly. And with improper or incomplete classification, the document may not show up in searches or cannot be found resulting in business/financial losses. In fact, a specialist should be able to tag a report as being associated with the game even when the word "cricket" is not used in the report – simply by imputing that when specific words/terms associated with the game are used, they refer to cricket.

Using automation to scale classification

Like cricket, every industry, from oil and gas to construction, from healthcare to fashion and from retail to entertainment has its lexicon and specialized terminology. As is most likely, 90% of any document will consist of easily-recognized terms; only 10% could be new terms. Can a machine learn the business terminology related to an industry and automatically, flawlessly and quickly begin to do the job of annotating, tagging and classifying documents with increasing specificity and organizing them into relevant classes?

Using automation to scale classification of documents has been around for some time. People have been turning to automation using natural language processing and machine learning to classify millions of documents accurately in a matter of minutes. What remains to be explored is hierarchical document classification, a science that involves forming a classification tree using the sub-categories to model real-world classes.

Consider the following example: reuters, the news organization, wants to process financial search queries for its customers. The organization has thousands, perhaps even millions, of articles it can retrieve based on a query like "2017 production data for automobiles". The customer would possible need to plough through hundreds of documents to find what they are looking for. Hierarchical document classification models the articles in such a way that all the articles related to automobile manufacturing are in some sense "children" of all news articles related to manufacturing. This in turn, are children of news articles related to trucks, cars, vans, trailers, recreational vehicles, motorcycles and scooters; these in turn are children of articles on diesel vehicles, petrol vehicles and electricity-powered and

solar-powered vehicles. We could go deeper or branch out in the tree with axels, gear box, spark plugs and drain bolts.
Often, humans – even subject matter experts—would have to read through more than half the document to confidently say that diesel cars were being referenced and not petrol driven vehicles. This is because most of the words in the document would be the same except for a few. Now, admittedly, humans are good at identifying these things; a subject matter expert (SME) would know that a diesel engine doesn't need a spark plug because the fuel is ignited by the heat of compression. The SME would, therefore, classify the document under "diesel" even though the document may not have used that word even once.
But human beings are slow. They would soon be overwhelmed by the rate at which documents are being generated and need to be classified.

What machine learning can do

This is where state-of-the-art Machine Learning (ML) steps in. It is a learning technique that uses algorithms to establish the relationship between words with zero human intervention. ML provides a natural way of annotating documents and classifying them into directories. When such a system classifies documents, an automobile engineer need not have to spend months looking into all documents on automobiles when searching for a document on a diesel engine's combustion chamber.
Machine learning and algorithms used for hierarchical document classification can be used in several situations. These methods can tell the difference between the syntactic and semantic similarities of words, can understand synonyms and even sarcasm, cleverly interpreting words that are context dependent.

Consider another example where an e-commerce company has millions of products. Tagging the products is a very important step so that there is efficient retrieval. Hierarchical document classification solves the problem of tagging in a flash. It classifies a new product, say a watch, into the right class. It then tags it as "Accessories" -> "Clocks" -> "Watch" -> "Brand Name", as it has already learnt all these relations using ML.
Another example could be of a pharma organization which typically has biomedical and clinical documents for lab tests, animal tests, comparison group studies, beta-phase, treatment modalities, acceptance by the drug regulatory authorities, etc for each product. They are likely to have millions of such documents. Hierarchical classification understands the relationship between these documents just by looking up the tree learned by the technique. Millions of complex words in documents are transformed into a picture of relationships.
Hierarchical document classification has several advantages allowing documents to be searched at various levels of topic specificities. And because it allows documents to be classified across classes, they can quickly and accurately access related topics.

About the authors:

Deependra Katiyar is a Practice Manager leading initiatives in the Enterprise Content Management (ECM) domain, with 17 years of IT experience spanning areas like ECM, customer experience, business process management, application modernization, integration and artificial intelligence. His major focus has been to identify and align the ECM offering to long-term market needs. He has successfully delivered large transformation programs across industry domains. He can be reached at deependra.katiyar@wipro.com.

Ameet Deshpande is interned at Wipro. He is a computer science student at Indian Institute of Technology, Madras. He specializes in problems relating to machine learning, artificial intelligence, and natural language processing, which lie at the intersection of industry and research. Along with Deependra, he worked on several projects and research in the area of enterprise content management.

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035, India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 160,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**

IND/BRD/DEC 2017-NOV 2018