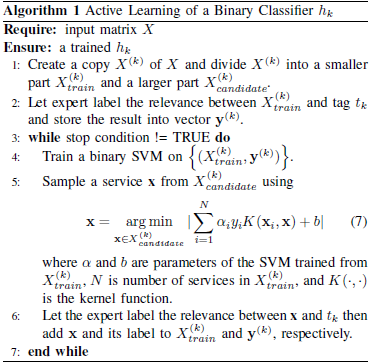
“Correlation-Aware Multi-Label Active Learning for Web Service Tag Recommendation” (We will use WSTB as a short abbreviation in the following content) is one of the selected research paper in the preparation phrase. The WSTB paper is published in 2017 IEEE 24th International Conference on Web Service written by the Computer Science professors in Rochester Institute of Technology (RIT). The major research of this research paper focuses on efficient tagging and accurate tag recommendation for web service. The propose approach uses a multi-label active learning workflow to identify the label correlation based on a small amount of high quality tagged web service provided by domain experts.

The introduction of the WSTB paper has raised a basic question about the disadvantage of traditional service discovery in both community and online equerry based platform. The traditional service discovery system heavily relies on the server developer’s effort on providing high quality service description in order categorize web services with accurate label. To reduce the huge tagging effort in a large-scale service pool, the content-based tag recommendation has helped is invented to extract related text information to identify the candidate tag. For better performance, researcher has followed a supervised learning strategy to predict the label of new input service based on the training dataset. However, the tag classifier requires high quality training dataset that has a very low miss-tagged rate. To address this problem, the multi-label active learning (AL) for web service can effectively filter out non-relevant tag from a small sized but high quality training dataset. AL uses an iterative approach to perform binary classification to determine correlation relation between difference input services and tags.

The “Tag Recommendation” section of the WSTB paper presents proposed AL with detail in math notations, tag correlation computation, actual algorithm description and tagging effort reduction. The setting paragraph introduces couple definition of input data including their math symbols. Those inputs are web service matrix linked with term frequency inverse document frequency (TF-IDF) score, tag matrix and correlation matrix between tags. The following part after the math notation description is the tag correlation computation. There are two tag correlation computation methods mentioned in this part: Jaccard clustering and Hierarchical clustering. The Jaccard similarity test measures the similarity between two sets of data with a range from 0% to 100%. The calculation of Jaccard index is straight forward. In the web service tagging recommendation scenario, the computer program identify shared number of web services and total number of web services between two set. Each set of web service has at least one common tag for its web service set. The next step is to divide the number of shared service by the total number of unique service in both set. If both set share no service, the computed result is 0% similar. In fact, the Jaccard clustering method is sensitive to small sample size or data sets with missing observations. The alternative tag method is Hierarchical clustering. This is an algorithm that groups similar tags into small clusters based on the closest distance. The following step re-computes the new distance between the new cluster and other tags/clusters. The algorithm repeat this process until there is only one cluster left. The linkage value for the last two clusters is the actual correlation.

Full implementation detail of Multi-Label Active Learning for Tag Recommendation is addressed in C paragraphs. With given high confident human labeled service information, the classification model generates a base classifier. This model will continuously consume input sample data to increase the labeling pool until the model reaches a certain accuracy with a validation dataset. This process is named as Active Learning of a Binary Classifier which is used to predict the relevance between all service and base modeling tag. The starting stage of the algorithm divides a full set of services into training and candidate categories. For the next step, web service domain expert classifies a small sized training dataset. Support Vector Machines (SVMs) is also part of the binary classification for base model training purpose. The objective of the SVMs is to find a hyperplane base on the training dataset and predefined tags provided by expert including the service description TF-IDF score that distinctly classifies the new sample service. The domain expert repeatedly examined new sample service from candidate set until the Binary SVM classifier reaches a predefined classification accuracy over a validation set.



The proposed correlation-aware active learning is defined in the very last section of the Tag Recommendation. This section gives clear analysis about web service tagging considering the comprehension time (CT) and decision time (DT). CT may take a lot longer time compared to the DT based on the length of service description. To tackle this problem, the system can significantly reduce the tagging process by merging multiple inform query into a single request with other relevant tag. For instance, one classifiers sample a service from the candidate set and predicts to assign the current tag. The application can unite other highly relevant tags to the current tag so the domain expert can check them in parallel. The tag correlation computation requires either Jaccard clustering or Hierarchical clustering.