**Problem Statement**

* The problem of diversifying keyword search is firstly studied in IR community. Most of them perform diversification as a post-processing or re-ranking step of document retrieval based on the analysis of result set and/or the query logs. In IR, keyword search diversification is designed at the topic or document level.
* Liu et al. is the first work to measure the difference of XML keyword search results by comparing their feature sets. However, the selection of feature set is limited to metadata in XML and it is also a method of post-process search result analysis.
* When the given keyword query only contains a small number of vague keywords, it would become a very challenging problem to derive the user’s search intention due to the high ambiguity of this type of keyword queries.
* Although sometimes user involvement is helpful to identify search intentions of keyword queries, a user’s interactive process may be time-consuming when the size of relevant result set is large.
* It is not always easy to get these useful taxonomy and query logs. In addition, the diversified results in IR are often modeled at document levels.
* A large number of structured XML queries may be generated and evaluated.
* There is no guarantee that the structured queries to be evaluated can find matched results due to the structural constraints;
* The process of constructing structured queries has to rely on the metadata information in XML data.

**Future Enhancement**

In order for a diversified search over plain xml source, we can improve the dataset architecture with multimedia contents specifically image files. Instead of plain contents, we can load styled contents so as to view it in an efficient manner. For that, We propose a technique to align data units into different groups so that the data units inside the same group have the same semantic. Instead of using only the DOM tree or other HTML tag tree structures of the SRRs to align the data units (like most current methods do), our approach also considers other important features shared among data units, such as their data types (DT), data contents (DC), presentation styles (PS), and adjacency (AD) information.