Evaluation Metrics for Information Access

A Tutorial at SEPLN 2014, Girona

Abstract

Evaluation metrics are not merely a tool to assess and compare systems. In the space of solutions to a problem, they work like a GPS that tells researchers where is the final destination, providing the operational definition of what systems should do. Remarkably, IR researchers can choose among a set of over one hundred metrics, all pointing at different places in the map, and in general there is no clear procedure to choose the most adequate metric in a specific scenario. And a wrong choice may imply falling off a cliff.

In this tutorial we will review and compare the most popular evaluation metrics for some of the most salient problems in Information Access, covering two basic types of task: text organization problems (Retrieval, Clustering, Filtering) and text generation problems (Summarization and Machine Translation). We will also discuss metric combination procedures.

The tutorial will make a special emphasis on the specification of formal constraints for suitable metrics in each of the tasks, and on the systematic comparison of metrics according to how they satisfy such constraints. This comparison provides criteria to select the most adequate metric or set of metrics for each specific Information Access task. The last part of the tutorial will investigate the grand challenge of providing a unified view of evaluation metrics for document organization tasks.

Tutorial Content

In this tutorial we will review and compare the most popular evaluation metrics for some of the most salient problems in Information Access, covering two basic types of task: text organization problems (Retrieval, Clustering, Filtering) and text generation problems (Summarization and Machine Translation). We will also discuss metric combination procedures.

For many Information Access problems, there are many competing evaluation metrics in the literature, and in general there is no clear procedure to choose the most adequate in a specific task/scenario. In practice, the tendency is often to choose the most popular metric (which has a snowball effect that tends to prefer the oldest

metrics). We cannot exclude the temptation for researchers to choose, among all available metrics, those that help corroborating their claims, or even to design a new metric to this aim. In addition, for many problems there are different quality aspects that are captured by different metrics (e.g. Precision and Recall) and, although Van Rijsbergen's F measure provides a way of combining and assigning relative weights to individual metrics, there is often a lack of clear criteria to assign relative weights between metrics. The need for metrics combination is of particular importance in text generation problems (summarization, translation), where there are many different criteria to compare system outputs with gold standard texts.

We believe that a better understanding of metrics, and of their conceptual, foundational, and formal properties, would help to avoid wasting time in tuning retrieval systems according to effectiveness metrics inadequate to specific purposes, and it will also induce researchers to make explicit and clarify the assumptions behind metrics.

The tutorial will make a special emphasis on the specification of formal constraints for suitable metrics in each of the tasks, and on the systematic comparison of metrics according to how they satisfy such constraints. This comparison provides criteria to select the most adequate metric or set of metrics for each specific Information Access task. The last part of the tutorial will investigate the grand challenge of providing a unified view of evaluation metrics for document organization tasks.

Structure

1) Evaluation Metrics for Document Organization Tasks

1.1 Clustering, Filtering, Retrieval

For each of these tasks we will analyze the formal restrictions that a suitable metric should satisfy, focusing on intuitive constraints on simple boundary conditions. According to how they satisfy the constraints, state of the art metrics

will be classified in families. We will also study cases in which metrics can be adapted to satisfy some of the formal constraints.

1.2 Metric combination

In most cases, there are two metrics to be combined (variants of precision and recall), and their relative weights may substantially influence evaluation results. We will study the properties of Van Rijsbergen's F measure (as the preferred metric combination function) and a method to measure the robustness of a result with respect to changes in the relative weighting chosen (the Unanimous Improvement Ratio).

1.3 Mixed problems

We will discuss practical problems where systems have to cluster and prioritize documents simultaneously, and discuss suitable evaluation metrics for these generalized "document organization" problem.

2) Evaluation Metrics for Generative Tasks

We will study evaluation metrics for Machine Translation and Text Summarization, which, in general, estimate the similarity between system outputs (peers) and human references (models).

We will focus on the basic properties of evaluation metrics which are based on text similarity, provide an exhaustive inventory of evaluation measures in the literature, discuss meta-evaluation procedures, and propose measure combination procedures.

Intended Audience

The tutorial contains material suitable both for novices and experts, but it is probably better classified as "advanced". Familiarization with Information Retrieval, Filtering, and Clustering is recommended.

Instructor(s)

Enrique Amigó (enrique@lsi.uned.es, UNED, Madrid, Spain) is associate professor at UNED and member of the nlp.uned.es research group. He has published several papers (in venues such as SIGIR, ACL, EMNLP, Journal of Artificial Intelligence Research, Information Retrieval journal, etc.) on evaluation methodologies and metrics for Text Summarization, Machine Translation, Text Clustering, Document Filtering, etc.

Publications: http://scholar.google.com/scholar?hl=en&q=enrique+amigo

Julio Gonzalo (julio@lsi.uned.es, UNED, Madrid, Spain) is head of nlp.uned.es, the UNED research group in Natural Language Processing and Information Retrieval. He has recently been CLEF 2011 General Co-Chair, Area Chair for EACL 2012, ECIR 2012 and EMNLP 2010, and co-organizer of the RepLab 2012/2013 Evaluation Campaigns for Online Reputation Management Systems and the WePS evaluation campaign on Web People Search systems. His research interests include Cross-Language and Interactive Information Retrieval, Search Results Organization, Entity-Oriented and Semantic Search, and Evaluation Methodologies and Metrics in Information Access.

Publications: http://scholar.google.com/citations?user=opFCmpYAAAAJ&hl=en

Both instructors have recently received, together with Stefano Mizzaro (U. Udine) a Google Faculty Research Award to pursue their work on Information Retrieval Evaluation Metrics.