

A Distributed Retrieval System for NTCIR-5 Patent Retrieval Task

Hiroki Tanioka Kenichi Yamamoto
Justsystem Corporation
Brains Park Tokushima-shi, Tokushima 771-0189, Japan
{hiroki_tanioka, kenichi_yamamoto}@justsystem.co.jp

Abstract

We developed a distributed search system with the corresponding very large scale corpora from NTCIR-5 Patent Retrieval Task. And we developed the method of query refining using Support Vector Machines. Our search system, which consists of 5 PCs could make indices of all claims for ten years. Additionally, we confirmed that our arranging the scoring method made an improvement of mean average precision.

Keywords: *distributed information retrieval, support vector machines, vector space model, inverted file*

1 Introduction

Our purposes to participate in NTCIR-5 Patent Retrieval Task is as follows.

- Research and development search systems which are corresponding a very large scale corpora.
- Research and development query refining methods which are useful for “invalidity search”

The background of first purpose is that digital documents are increasing in recent years, while we need search systems to effectively access these documents. But traditional search systems cannot make the full text indexes for these documents. Therefore we propose a distributed search system which is build on a distributed framework.

A background of second purpose is that it is high cost to make query from claims manually for invalidity patent search. Then, we try to make query from claims automatically.

The rest of this paper is divided into three sections. Section 2, we describe an architecture of our distributed processing framework and search system. Section 3, we describe results of formal runs. Section 4, we discuss about results and future works.

2 System Description

In this section we describe the architecture of our distributed search system and information retrieval models including some scoring methods.

2.1 Document Retrieval Subtask

First, we explain the system architecture and models for Document Retrieval Subtask.

2.1.1 Overview

We develop a distributed search system which is based on Vector Space Model using term partitioning with an inverted file-based system, while a single inverted file is created for the document collection and the inverted lists are spread across the processors.

During query evaluation, the query is decomposed into indexing items and each indexing item is sent to the processor that holds the corresponding inverted list[1].

2.1.2 Distributed Processing Framework

Cocktail Framework¹ is used to make the distributed search system based on Vector Space Model. The framework provides a service of agents between client and server as broker between query and each indexing item.

Figure 1 shows an overview of this framework. To process a job² in our system, a client machine receive a job, and keep in a FIFO queue. And then, to send the job to a server machine, unconfined agents pull the job from the FIFO queue. Last, the server machine performs the job and send a result back to the client machine via same agent.

¹Cocktail Framework is developed for distributed processing framework by Justsystem Corporation.

²Job is described as a pair of command and argument which are processed in our system.

