Resumen "Servicios de cache distribuidos para motores de búsqueda web"
Carlos Gómez

Los Motores de Búsqueda Web (WSEs) actuales están formados por cientos de nodos de procesamiento, los cuales están particionados en grupos llamados servicios. Cada servicio lleva a cabo una función específica, entre los que se destacan: (i) Servicio de Front-End; (ii) Servicio de Cache; y (iii) Servicio de Índice. Específicamente, el Servicio de Front-End maneja las consultas de usuario que arriban al WSE, las distribuye entre los otros servicios, espera por los resultados y genera la respuesta final al usuario. La idea clave del Servicio de Cache es reutilizar resultados previamente computados a consultas hechas en el pasado, lo cual reduce la utilización de recursos y las latencias asociadas. Finalmente, el Servicio de Índice utiliza un índice invertido para obtener de manera eficiente los identificadores de documentos que mejor responden la consulta. El presente trabajo de tesis se focaliza en el diseño e implementación de servicios de cache distribuidos eficientes.

Varios aspectos del sistema y el tráfico de consultas deben ser considerados en el diseño de servicios de cache eficientes: (i) distribuciones sesgadas de las consultas de usuario; (ii) nodos que entran y salen de los servicios (de una forma planificada o súbitamente); y (iii) la aparición de consultas en ráfaga. Cualquiera de estos tópicos es un problema importante, ya que (i) genera una asignación de carga desbalanceada entre los nodos; el tópico (ii) impacta en el servicio cuando no se utilizan mecanismos de balance de carga dinámicos, empeorando la asignación desbalanceada de carga y perdiendo información importante ante fallas; y finalmente (iii) puede congestionar o dejar fuera de servicio algunos nodos debido al abrupto incremento en el tráfico experimentado, incluso si se tiene un servicio balanceado. Dada la arquitectura que se emplea en este trabajo, el Servicio de Cache es el más expuesto a los problemas mencionados, poniendo en riesgo la tasa de hit de este servicio clave y el tiempo de respuesta del WSE.

Este trabajo ataca los problemas mencionados anteriormente proponiendo mejoras arquitecturales, tales como un enfoque de balance de carga dinámico para servicios de cache altamente acoplados (desplegados en clusters) basados en Consistent Hashing, y un esquema para monitoreo y distribución de consultas frecuentes. El mecanismo de balance de carga propuesto es una nueva solución al problema de balance de carga en clusters de computadores que corren aplicaciones manejadas por los datos. Además, se estudia cómo predecir la aparición de consultas en ráfaga para tomar acciones correctivas antes de que saturen o colapsen algunos nodos. Finalmente, se adopta la idea de un sistema tolerante a fallas para proteger información valiosa obtenida a través del tiempo. La idea fundamental es replicar algunas entradas de cache entre distintos nodos para que sean usados en caso de fallas.