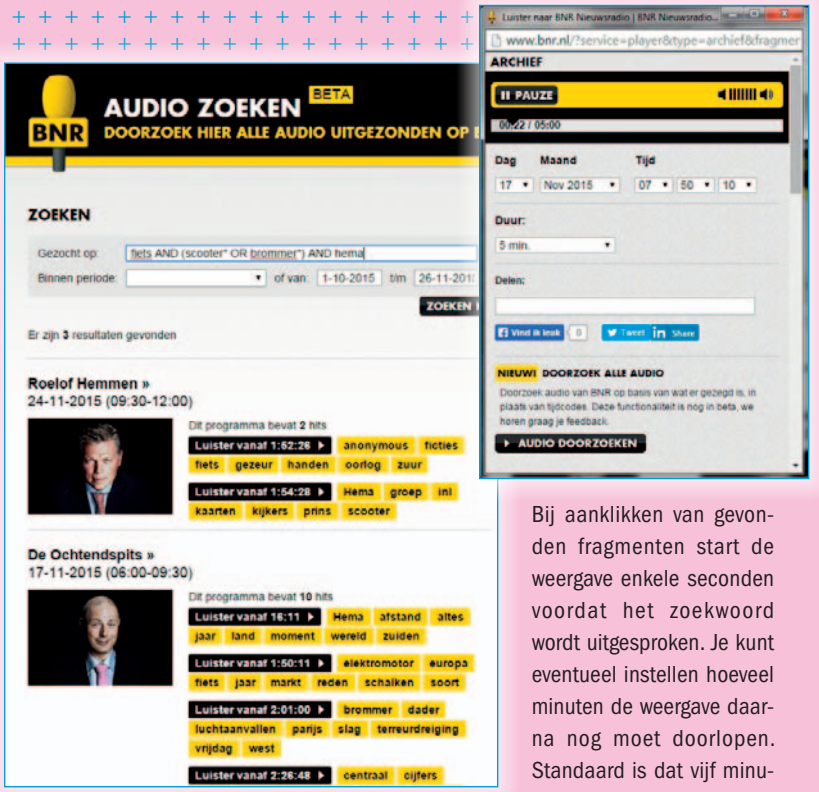
DAT ZOEKEN WE OP

Audio zoeken

IP bekijkt dit keer niet de zoekfunctionaliteit van één specifieke site, maar verdiept zich in het full-text zoeken in één specifieke soort materiaal: gesproken woord.

Door: Eric Sieverts



Wie op ‘audio zoeken’ zoekt, vindt vooral verwijzingen naar systemen die ingesproken tekst omzetten in opdrachten voor een gewoon zoeksysteem. Met de komst van mobiele apparatuur heeft dat een grote vlucht genomen. Inspreken gaat daar veel makkelijker dan intoetsen op een pietepeuterig touchscreen. Maar over die toepassing wil ik het niet hebben. Mij gaat het om het omgekeerde. Hoe je via het toetsenbord kunt zoeken in tekst die wordt uitgesproken in audio- en videobestanden. De techniek van spraakherkenning is dankzij de eerstgenoemde toepassingen al aardig goed – en niet meer alleen voor Engels. Omzetten van gesproken woord naar met gewone zoekmachines doorzoekbare tekst is in principe dus al goed mogelijk. Wie een zoekmachine koopt om eigen digitale informatie doorzoekbaar te maken, kan daarbij vaak een audiozoekmachine meegeleverd krijgen, om bijvoorbeeld opgenomen telefoongesprekken te kunnen terugvinden op hun inhoud in plaats van op metadata. Maar algemene audio-googles voor het web bestaan nog niet.

Wat je op het web wel tegenkomt, zijn óf demo’s van leveranciers van de eerdergenoemde enterprise search-software, óf zoeksystemen voor heel specifieke collecties. Tot deze aflevering werd ik dan ook geïnspireerd door BNR Nieuwsradio die sinds kort zo’n zoekdienst aanbiedt. Van zo’n afgeperkte collectie weet je in elk geval wat erin zit. Bij Voxalead en Audiosear.ch, voorbeelden van demosystemen die hier aan de orde komen, weet je dat eigenlijk niet. Deze drie diensten hebben wel gemeen dat wat je ermee vindt, in geen geval op dezelfde manier ook met Google te vinden is. Een belangrijke functionaliteit van alle drie is, dat ze je meteen naar het fragment brengen waar je zoekwoorden worden uitgesproken, zodat je daarvoor niet een hele uitzending van een uur hoeft af te luisteren.

BNR Nieuwsradio – audiozoeken.bnr.nl

Zelf noemen ze het nog een bèta-versie, maar het werkt eigenlijk al verbazend goed. Je kunt zoekwoorden Booleaans combineren, met AND, OR, NOT en haakjes en je kunt ook trunceren. Dat combineren werkt dan binnen de tekst van een hele uitzending van vaak meer dan twee of drie uur, waarin een veelheid aan onderwerpen de revue passeert.

In het zoekresultaat zie je welke fragmenten in welke uitzendingen zijn gevonden. Bij elk fragment zie je welk van je zoekwoorden erin voorkomt. Bovendien wordt een aantal andere woorden getoond die in de buurt van dat zoekwoord werden uitgesproken. Dat helpt aardig bij het inschatten van de relevantie van het gevonden fragment.

Bij aanklikken van gevonden fragmenten start de weergave enkele seconden voordat het zoekwoord wordt uitgesproken. Je kunt eventueel instellen hoeveel minuten de weergave daarna nog moet doorlopen. Standaard is dat vijf minuten, wat vaak al te lang is.

Voxalead – voxalead.labs.exalead.com/search

Dit is een demo voor de audiofunctionaliteit van Exalead search. Er zit spraakherkenning in voor negen talen, waaronder Nederlands. Die herkenning werkt aardig goed (al bleek het systeem het gebrabbel van Jan Peter Balkenende al net zo slecht te verstaan als ikzelf). Je kunt dat mooi zien en controleren door het transcript mee te laten lopen bij het afspelen van gevonden audio- of videofragmenten. Helaas zit er geen nieuwer materiaal in dan uit januari 2014. Maar het is zeker leuk om eens uit te proberen. En dat is natuurlijk net de bedoeling van zo’n demo.

Audiosear.ch – www.audiosear.ch

Ze noemen zich ‘A full text search & recommendation API for podcasts and radio’. Er zit alleen Engels materiaal in; een nog steeds groeiende collectie van meer dan 450 podcasts en radioprogramma’s, het meeste via iTunes. Ook hier wordt mooi getoond waar de zoekwoorden worden uitgesproken met daarbij een stukje transcriptie van de tekst; bij het afluisteren kun je die ook laten meelopen. Deze dienst biedt zelfs een alertfunctie die je erop attendeert als jouw zoektermen in nieuwe uitzendingen worden uitgesproken. <

Eric Sieverts, redacteur van IP en freelance docent en adviseur